Identification and characterisation of the Cdx1 and Apc1 cis-regulatory elements in mouse and Fugu rubripes

Juarez-Morales, Jose-Luis

Award date:
2005

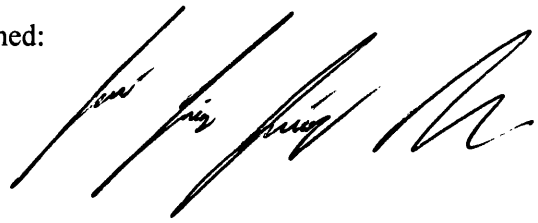
Awarding institution:
University of Bath

Identification and characterisation of the *Cdx1* and *Apc1* cis- regulatory elements in mouse and *Fugu rubripes*

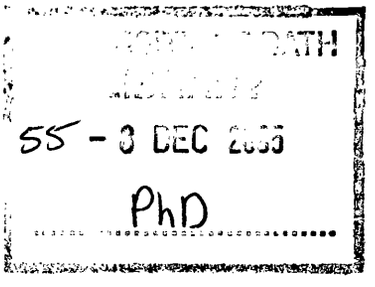
Submitted by **José-Luis Juárez-Morales**

For the degree of Doctor of Philosophy at the University of Bath, 2005.

Signed:



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346


55 - 3 DEC 2003
PhD

Abstract

The *Cdx* homeobox transcription factor family are expressed during development regulating axial skeletal development. Later, *Cdx1* and *Cdx2* expression is restricted to the crypt–villus axis in the differentiating intestinal epithelium. However, the regulatory mechanisms governing *Cdx1* are not completly understood. In this study, the *Cdx* genes of *Fugu rubripes* were identified and characterized. Consistently, the *Fugu Cdx* family contains the same gene structure and tissue distribution as their mammalian orthologs. To investigate the regulatory mechanisms controlling the *Cdx1* expression reporter mouse *Cdx1 LacZ* and *Fugu Cdx1* GFP constructs were assayed in intestinal cells. Cis-regulatory elements located upstream of the mouse and *Fugu Cdx1* were able, and even enhanced, the expression of the reporter in the cells.

Subsequently, functional expression assays in the zebrafish embryo, revealed that the *Fugu Cdx1* upstream region directs the expression of the transgene during gastrulation, and at 14-somite stage, in the posterior region of the embryo. Further comparative analysis of the *Fugu*, mouse and human 5' flanking regions showed conserved transcription factor binding sites across species.

The Adenomatous Polyposis Coli (Apc) multifunctional protein is involved in cell adhesion, cell migration and cell proliferation. In tissues like the intestine, Apc is involved in the regulation of β-catenin in the Wnt signalling pathway. When β-catenin translocates into the nucleus, it forms a complex with the Tcf/Lef transcription factors to activate genes involved in proliferation and differentiation of the cells (e.g. *Cdx1*). When the Wnt pathway is activated, an inhibitory complex formed by Axin, GSK3β and Apc down-regulates β-catenin; transcriptional activation of genes does not take place. Apc regulates indirectly the proliferation and differentiation changes in the intestinal cells.

To investigate *Apc1* regulation and its potential interaction with the Cdx proteins, comparative studies to look for conserved regulatory regions and transcription factors binding sites were performed using the *Fugu rubripes* genome.

We identified and characterized the *Fugu Apc1* and *Apc2* genes. Comparative studies showed a good level of conservation of the genomic, gene and amino acid structures of the *Apc* genes between mouse, human and *Fugu*. RT-PCR analyses revealed that the *Fugu Apc* genes are expressed, in the same variety of tissues and during development as the mouse orthologs.

Comparative analyses of the mouse and *Fugu Apc1* upstream sequences showed that, even though their sequences have diverged during evolution, transcription factor binding sites (e.g. Cdx, GATA and HNF) remain conserved between the species. Transfection and transactivation assays using the mouse *Apc1* upstream sequence and the Cdx1 and Cdx2 cDNA, suggest an antagonistic role of the Caudal proteins in the regulation of *mApc1* in intestinal cells.

The expression distribution of the *Fugu Apc1* gene, and the presence of conserved transcription factor binding sites between mouse and *Fugu* upstream regions, suggest that the regulatory mechanism of the *Apc1* in the intestinal tissue are conserved between species.

Acknowledgements

I thank Dr. V. Subramanian for the space, supervision and helpful scientific discussions during the PhD. I thank Dr. U. Strahle for the pCS2:GFP and pCS2:LacZ vectors, Dr. A. Chandrasekhar for the preliminary data with *frCdx1* GFP expression in the zebrafish embryo and Dr. L. Colletta for the p*mApc* CAT clone.

Special thanks to Dr. Greg Elgar for the space and time provided during my stay in the HGMP-MRC Cambridge. I would also like to thank the Comparative Genomics group; Mel, Ivonne, Debbie, Phil and Tanya for their support and for making me feel one of their own group.

I am grateful to CONACyT-Mexico for the scholarship that allowed me to do my PhD studies in the UK. To the University of Bath for providing the space and environment to accomplish my studies.

Thank you to Professor Stuart Reynolds and Dr. Michael Wride for their useful comments and advice toward this work.

Thank you to the Lab 0.73, previous and present members; Jason, Karolina, Ben, Richard, Mervyn and Abhishek for all the help, jokes and great moments that got me through my PhD.

Thanks to all the people who kept me sane, healthy and rolling during my time in the UK; Cesare, John, Gloria, Diya, Kim (for moral and material support), to Michael and Waqas in Cambridge for their friendship and good memories.

I am deeply grateful to Professor Robert Eisenthal for his time, patience, help and understanding in the most difficult moments of my PhD.

Finally, Gracias a José Luis Juárez, Carmen Morales y Ana Luisa Juárez Morales por estar ahí, por las palabras de aliento, por el apoyo brindado, en éste y en todos los proyectos que me he trazado en mi vida.

A Elena Mikel por su paciencia, apoyo, cuidado, alegría, por todas las risas y ayuda que hicieron posible el completar esta tesis de doctorado.

Table of Contents

List of Abbreviations

Aa	Amino acid
A-P	Anterior posterior axis
Apc	Adenomatous polyposis coli
ARCS	Adapted related complex subunit gene
Asef	APC- stimulated guanine nucleotide exchange factor
Bcd	*Bicoid* gene
BLAST	Basic Local Alignment Search Tool
bp	Base pair
BSA	Bovine serum albumin
BSKS2+	Bluescript II SK+ vector
β-TRCP	β-transducin repeat-containing protein
Caco2	Caucasian colon carcinoma cells
Cad	*Drosophila Caudal* gene
CAMK2A	Calcium/calmodulin-dependent protein kinase type I alpha chain gene
Carr	Arrestin β2 red cell isoform gene
CAT	Chloramphenicol acetyl transferase
CIP	Calf Intestinal Phosphatase
CMV	Cytomegalovirus promoter
Dkk1	*Dickkopf1* gene
Dazap1	Daz associated protein 1
DMEM	Dulbeccos Modified Eagle Media
DMSO	Dimethyl sulphoxide
DNA	Deoxyribonucleic acid
EDTA	Ethylenediamine tetra-acetic acid
ELISA	Enzyme-linked immunosorbent assay
EMBL	European Molecular Biology Laboratory
EMSAs	Electro mobility shift assays
Est	Expressed- sequence tag
FAP	Familial Adenomatous Polyposis
Fk	*Fukutin* gene
FLK1	Tyrosine kinase receptor gene

Flt3	Cytokine receptor precursor gene
GFP	Green fluorescence protein
GPD	Glyceraldehyde-3-phosphatase dehydrogenase gene
Hb	*Hunchback* gene
HDLG	Human disc large protein
HMG	High mobility group
Hpf	Hours post fertilisation
Hox	Homeobox genes
ICM	Inner cell mass
IEC-6	Normal rat epithelial cells
I-FABP	Intestinal type fatty acid- binding protein
IP	Insulin precursor protein gene
Ipf1	Insulin promoter factor gene
Kb	Kilo base
LB	Luria broth
LEF	Lymphoid enhancer factor
Lnx2	Zinc finger family gene
Mb	Mega base
Mcsfr	Macrophage colony stimulator factor I receptor precursor
$MgCl_2$	Magnesium chloride
MGI	Mouse Genome Bioinformatics
NaCl	Sodium chloride
NCBI	National Centre for Biotechnology Information
NIX	Nucleotide identification program
Ntl	*No tail* gene
OD	Optical density
OMIM	Online Mendelian Inheritance in Man
PAPI	Pancreatitis associated protein I
PCSK4	Convertase subtisilin kekintype 4 gene
PBS	Phosphate buffered saline
Pdgfrβ	Beta platelet- derived growth factor receptor precursor gene
PEG	Polyethylene glycol
PPA2	Phosphatase 2A protein

RA	Retinoic Acid
RACE	Rapid amplification of cDNA ends
RAR	Retinoic acid receptor
RARE	Retinoic acid receptor element
RDHL	Retinol dehydrogenase L gene
RFP	Red fluorescence protein
RFP	Finger protein 26 gene
RT	Reverse transcriptase
RT-PCR	Reverse transcriptase polymerase chain reaction
Rsp15	Ribosomal protein S15
RXR	Retinoic X receptor
SAMP	Specific Axin binding domain
SDS	Sodium dodecyl sulphate
SI	Sucrose isomaltase gene
SLC6A7	Solute carrier family 6 gene
Spt	*Spadetail* gene
TAE	Tris–acetate EDTA
TBE	Tcf binding elements
TBE	Tris borate EDTA
TCF	Human T cell factor
TE	Tris- EDTA
TESS	Transcription element search system
TFBS	Transcription factor binding site
TFSEARCH	Searching transcription factor binding sites program
Tk	Thymidine kinase promoter
TNE	Tris and Sodium buffer EDTA
Tris-base	Tris- hydroxymethyl methylamine
TrK-A	Slow nerve growth factor receptor gene
TSS	Transcription start site
TXS	Testis express gene
UTR	Untranslated region
WR	Working reagent
X-gal	5-bromo-4-chloro-3-indolyl-β -galactopyranoside

Chapter One

Introduction

1.1 Anterior posterior axis formation

1.1.1 Formation of the embryo

The development of the embryo is a very well conserved process. The first stage of development occurs after the fertilized egg has undergone 4 cell divisions, at the 16 cell stage, the embryo is called morula, which consists of a group of inner cells surrounded by larger cells. The cells derived from the external cells will form the trophoblast cells. At this stage, the first two cell lineages of the embryo have originated, the trophectoderm and the inner cell mass. The trophectoderm gives rise to extraembryonic structures.

The trophoblast cells, also known as trophectoderm, do not contribute to the embryo proper. The trophectoderm will give rise to the extra embryonic structures (e.g. placenta) and the inner cell mass (ICM) will develop into the embryo proper. By the time of implantation, the embryo has developed to a vesicular structure called the blastocyst (32 cell stage). The blastocyst contains an external cell layer that begins to expand as fluid fills the internal cavity. The inner cell mass is attached to the internal side of the vesicle.

The inner cell mass divides into two, the primitive ectoderm and the primitive endoderm or epiblast. The primitive endoderm will contribute to the formation of the extra embryonic membranes such as the parietal endoderm and the visceral endoderm. The primitive ectoderm also contributes to the embryo and also some extra embryonic membranes.

After this stage, the embryo begins to acquire a cup shaped appearance and at 6.5 days after fertilization, the future axis of the embryo becomes visible with the onset of gastrulation and the formation of the primitive streak. The start of the primitive streak marks the posterior of the embryo. These processes are tightly regulated at the genetic level (Figure 1.1).

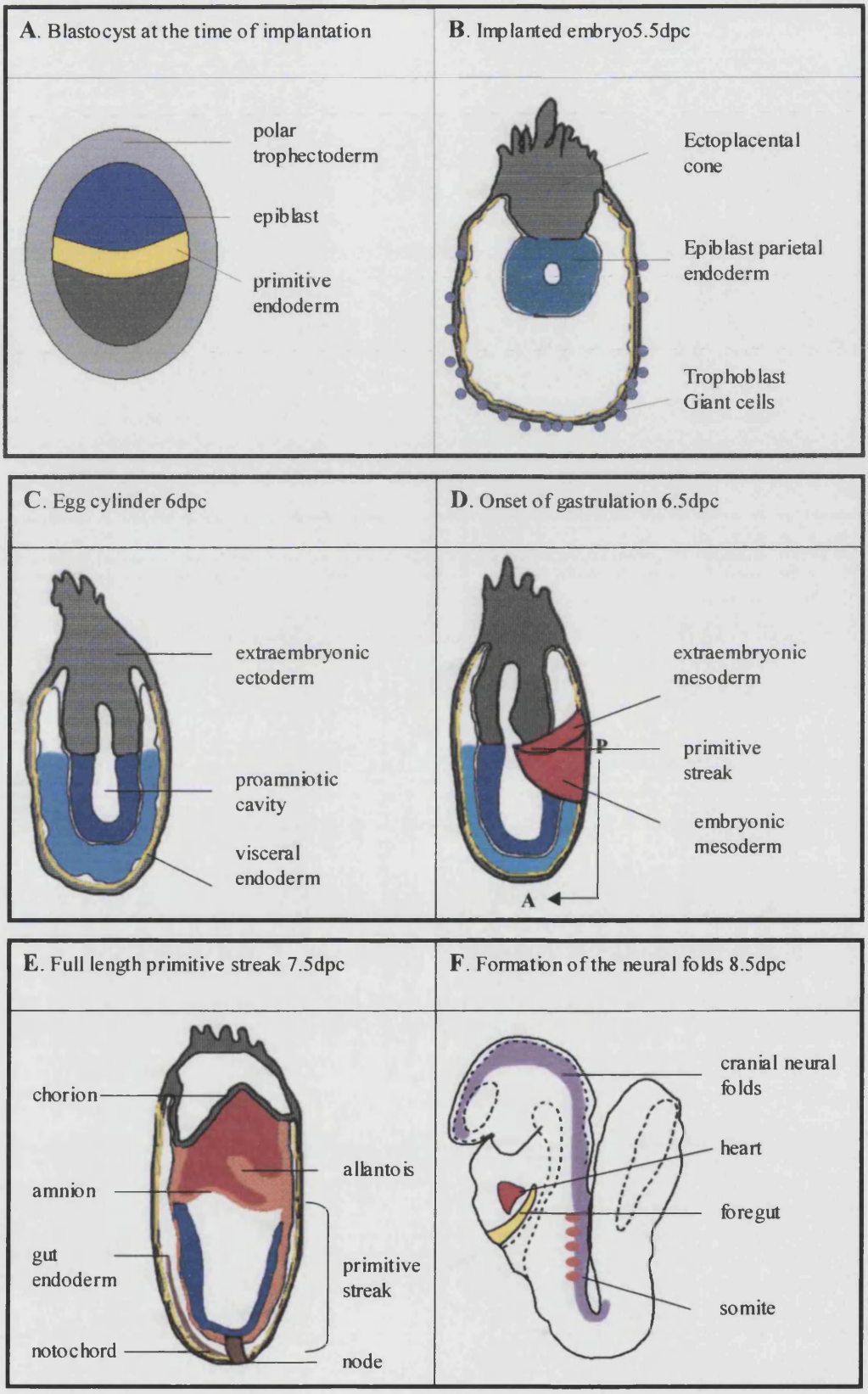


Figure. 1.1. Early development of the mouse embryo. **A.** After cleavage, the blastocyst forms the trophectoderm and the ICM. The ICM divides into the

primitive ectoderm or epiblast and the primitive endoderm. **B.** The polar trophectoderm in contact with the epiblast forms the extraembryonic tissues, the ectoplacental cone and mural trophectoderm. The epiblast elongates and forms a cup-shape structure. **C.** The epiblast and extraembryonic tissue form a cylindrical structure known as the egg cylinder. **D.** The primitive streak starts at the posterior of the epiblast (P) and extends interiorly (A). The epiblast cells migrating through the streak will become mesoderm and endoderm. **E.** The primitive streak continues its extension; the extraembryonic mesoderm develops at the posterior of the primitive streak. The visceral yolk sac, the allantois and chorion will be part of the placenta. **F.** At the end of gastrulation, organogenesis takes place with the appearance of heart, cranial neural folds and somites (Wolpert *et al.* 2002)

During gastrulation the cell sheets rearrange, occupy new positions and find new neighbours to generate the three main germ layers –ectoderm, mesoderm and endoderm- that will rise to all the tissues of the body. At the end of gastrulation, the heart, liver and gut have acquired their respective positions in the embryo; the head becomes distinct and the forelimb buds start to develop. Organogenesis follows and culminates in development of the adult with the creation of specific tissues and organs (Gilbert 2000; Wolpert *et al.* 2002).

1.1.2 Anterior posterior axis formation

The anterior posterior axis of the organism is established at a very early stage of development. As mentioned earlier, gastrulation marks the establisment of the anterior posterior axis (A- P axis) in the embryo. At 6.5 days the cells of the epiblast, a group of cells that will give rise to the embryo proper, start to form the primitive streak. The site of formation of the streak marks the posterior end of the embryo, and the cells from this region move towards the anterior part forming the primitive streak. In the anterior region some cells establish a signaling node, the organizer. *Goosecoid*, *Nodal*, *Lim-1* and *HNF3β* are genes expressed in the node and are involved in the initiation and maintenance of the primitive streak as well as in the formation of the most anterior regions of the embryo like the head.

Once the A- P axis has been established, the patterning and identity of the structures is specified by the *Hox* genes. These homeobox genes provide the cells with positional information to migrate and locate in the proper position to generate future structures. The *Hox* genes are expressed in mesodermal cells present in the primitive streak. The paraxial mesoderm then becomes segmented into somites that give rise to the muscles, skin and axial skeleton (Marshall *et al.* 1996).

Studies using different model systems (Conlon 1995; Gellon and McGinnis 1998) have shown the contribution of the *Hox* genes to the establishment and formation of the A- P axis in the developing embryo. These studies demonstrated that a shift or ablation of specific *Hox* genes produce alterations in somites and vertebral transformation.

However, specification of the A-P axis does not depend on the expression of the *Hox* genes. Studies have shown that the murine *Cdx* genes as well as its homolog *Caudal* in *Drosophila* are essential for the A-P axis formation during early development. *Cdx1* has been shown to bind to the *Hoxa-7* promoter, a gene involved is somitogenesis (Subramanian *et al.* 1995). The *Hoxb-8a* gene expressed in the central nervous system during development is also regulated by Cdx proteins, and ectopic expression of *Cdx* anteriorizes the expression of *Hoxb-8* in the embryo (Charite *et al.* 1998). The *Hoxc-8* is expressed in the embryonic tissues where *Cdx2* is also expressed, two specific *Cdx2* binding sites identified in the *Hoxc-8* upstream regulatory region have been shown to drive the expression of the gene; furthermore, these Cdx2 binding sites were able to activate heterologous promoters when placed in an enhancer position (Taylor *et al.* 1997).

1.1.3 Role of the *Cdx* in anterior-posterior axis formation

The *Caudal* related homeodomain transcription factors, *Cdx1*, *Cdx2* and *Cdx4*, are known to be expressed early during development and contribute to the A-P axis formation in the mouse embryo.

During development, *Cdx1* is expressed during late gastrulation at 7.5 dpc. *Cdx1* shows a gradient of expression, being the posterior end of the primitive streak more active in *Cdx1* expression than the anterior end. Paraxial mesoderm and neuroectoderm also show expression at this stage. By 8.5 dpc expression is found in

the neuroectoderm and in the mesoderm of the developing somites. Between 8.25 and 8.75dpc the most anterior expression of *Cdx1* is at the base of the hindbrain. This expression boundary then moves to the spinal cord limit. After the formation of the somites, the anterior boundary of expression is at the spinal cord and the dorsal region of the somites. At 9.0 dpc, the mesoderm begins to form the forelimb bud between the seventh and tenth somites. By 9.5 dpc *Cdx1* expression is located in the dorsal region of the anterior somites and in the mesodermal cells of the forelimb buds. After this stage, expression of the transcript decreases markedly, although the Cdx1 protein is still found until 12 dpc in the somites (Meyer and Gruss 1993).

In contrast to *Cdx1*, *Cdx2* is expressed much earlier at 3.5 dpc in the trophectoderm of the blastocyst. From 4.0 to 6.5 dpc, *Cdx2* expression is detected in extraembryonic ectoderm. At 7.5 dpc, expression is present in the chorion, the most external extraembryonic membranes involved in respiratory exchange; the mesoderm of the developing allantoic bud, the allantoids is a set of extraembryonic membranes where blood is carried in and out of the placenta; the ectoplacental cone, which is derived from the extraembryonic ectoderm and attaches to the uterine wall and posterior primitive streak.

At 8.5 dpc, *Cdx2* expression is seen in all three germ layers in the posterior region of the embryo; the anterior expression is found in the neural tube, neural plate and notochord, a mesodermal derived structure beneath the future central nervous system. The paraxial mesoderm is also positive in expression. By 9.5 dpc, the tail bud, neural tube, notochord posterior and gut endoderm of the embryo remains positive, although no more expression is seen anteriorly to the foregut/midgut junction. At this stage, there is a gradual decrease of expression in the notochord and neural plate.

At 12.5 dpc, expression is restricted to the gut endoderm; gut epithelium shows strong expression in the foregut/midgut boundary, and the expression decreases towards the posterior part of the hindgut. The epithelium is negative in expression in the rectum and urogenital sinus. At this stage, the cytotrophoblast, a structure derived from the trophectoderm is also positive for *Cdx2* expression (Beck *et al.* 1995).

The third Caudal related homebox factor, *Cdx4*, is also expressed during early stages in the posterior region of the developing embryo. The allantois and the posterior end of the primitive streak are positive for *Cdx4* expression at 7.0-7.5 dpc.

By 8.5 dpc, *Cdx4* is highly expressed in the posterior neural tube, presomitic mesoderm and hindgut endoderm. At this stage there is a gradual expression of the gene being the posterior part of the embryo the strongest in expression. The anterior limit of expression lies posterior to the youngest somite.

At 9.5 dpc, *Cdx4* expression is still present in the tail bud of the embryo, neural tube, early mesoderm and the epithelium of the hindgut are positive in expression, while no expression is found in any anterior structure. No expression has been detected after 10.5 dpc (Gamer and Wright 1993).

1.2 Intestinal crypt villus axis

1.2.1 Structure and histology of the intestine

In mammals including mouse, the formation of the gut begins after gastrulation when the endoderm invaginates to create the anterior and posterior intestinal portals. These ends elongate inwards and finally fuse to create a primitive intestinal tube surrounded by mesodermal tissue. This process happens at 8.5 dpc when the embryo turns its body. As development proceeds, three main regions start to differentiate in the gut tube: the foregut, from which will originate the pharynx, oesophagus and stomach; the midgut that will give rise to the small intestine; and the hindgut that gives rise to the large intestine. After birth, ingestion and primary digestion occur in the foregut; the completion of digestion and absorption of nutrients take place in the small intestine; and water and salt reabsortion as well as ejection of waste are performed in the large intestine (Kaufman *et al.* 1997).

The anatomy of the intestine is similar in the midgut and hindgut; both organs consist of epithelium, surrounded by connective tissue. These two tissues are surrounded by concentric layers of smooth tissue innervated by enteric nerves. The function of the enteric nerves is to allow contractions of the intestine during digestion (Kedinger *et al.* 1998).

The intestinal epithelium is a renewing tissue where proliferation, differentiation and migration of cells occur through the life of the individual in a tightly regulated process. There are two axes of differentiation in the intestine, the anterior posterior axis (A-P) and the crypt-villus axis (C-V). The crypt (also known as

crypts of Lieberkühn) is a proliferative compartment located at the bottom of the CV axis, where a reservoir of stem cells is located. The villus is the top compartment where differentiated cells migrate to take their place.

Bjerknes et al. (1999) proposed that two populations of cells are derived from each stem cell, a population that will give rise to the absorptive enterocytes and a population that will produce the different secretory cells (goblet, enteroendocrine and paneth). The Paneth cells are the only differentiated cell types that stay at the bottom of the crypt. The other three types, goblet cells, enteroendocrine cells and absorptive enterocytes differentiate as they migrate to the villus (Potten and Loeffler 1990; Stappenbeck *et al.* 2002).

1.2.2 Role of the Cdx in intestinal epithelial differentiation

The *Caudal* family, specifically *Cdx1* and *Cdx2* whose second phase of expression is restricted to the developing intestine and maintained through adulthood, are early players in the specification of the intestine (Duprey *et al.* 1988; James and Kazenwadel 1991; Beck *et al.* 1995; Subramanian *et al.* 1998)

As the intestine develops and matures, the A-P axis is distinguished during the epithelial differentiation along the cephalo-caudal regions, resulting in morphological and biochemical differences between the small and large intestine (16-17dpc). Once the cells are specified as gut primordia, a regional patterning marks the foregut, midgut and hindgut development derived from endoderm tissue. At this stage, expression of *Cdx1* and *Cdx2* is restricted to the intestinal epithelium. *Cdx2* is especially important for hindgut development (Freund *et al.* 2003). The *Cdx* genes may influence the A-P patterning through regulation of the *Hox* genes (Chawengsaksophak *et al.* 1997; Charite *et al.* 1998).

At 14.5 dpc, a weak expression of *Cdx1* reappears in the small and large intestine epithelium. At 17.5 dpc, the expression becomes stronger in both tissues (Duprey *et al.* 1988; Subramanian *et al.* 1998). The increase of this expression obeys to the developmental transition of the gut. Then, expression is restricted to the base of the primitive villi. Twenty days after birth, the finger-like villi separated by the crypts are a characteristic feature of the small intestine. In contrast, crypts are the main feature in the colon. *Cdx1* expression is located at the base of the crypts. Expression in the anterior posterior axis goes from the duodenum to the colon (Figure 1.2), the

expression being stronger in the colon and decreasing gradually in the anterior region (James and Kazenwadel 1991; Subramanian *et al.* 1998).

From 12.5 dpc onwards *Cdx2* is expressed in the epithelium of the gut; there is an anterior posterior pattern in expression. In the proximal colon transcripts are more abundant than in the distal colon. The expression of *Cdx2* in the proximal colon is stronger in undifferentiated cells at the base of the crypt than in differentiated cells at the top of the crypt. In contrast, in the distal colon expression is stronger in differentiated cells (Figure 1.2) situated in the top half of the crypt. (James *et al.* 1994; Silberg *et al.* 2000).

By 15 dpc, *Cdx2* is expressed in the endoderm intestinal epithelium (Figure 1.2), the expression continues in the adult intestinal epithelium, in the villus of the small intestine and in the crypts of the colon (James *et al.* 1994). Due to the expression pattern shown by *Cdx1* and *Cdx2* in the crypt-villus axis, *Cdx1* is essential in maintaining the population of cells that come from the stem cell niche, and cells that undergo differentiation. In contrast, the role of *Cdx2* in this axis seems connected with maturation and differentiation of cells situated outside of the crypt.

Cdx2 is expressed in the differentiated enterocytes and regulates the expression of enterocytic genes. Reports have shown that *Cdx2* is involved in the regulation of specific intestinal genes such as those encoding sucrase isomaltase, carbonic anhydrase 1, apoliprotein B, among others (Suh *et al.* 1994; Drummond *et al.* 1996; Lee *et al.* 1996).

The *Cdx2* binding sites have to be located in an enhancer context in order to show specific cell type activation (Taylor *et al.* 1997), suggesting the involvement of a co- activator factor to mediate the binding of *Cdx2* in its binding site. For example, *Cdx2* forms a transcriptional complex with *HNF-1α* and *GATA4* to activate the sucrose isomaltase (SI) gene in the epithelial cells in the villus (Boudreau *et al.* 2002).

Cdx1 has been found to activate the pancreatitis associated protein I promoter, a gene involved in cell proliferation in the crypts (Moucadel *et al.* 2001). Lynch et al (2000) reported that *Cdx1* is able to inhibit growth of intestinal epithelial cells by disrupting the G1 phase of the cell cycle. This inhibition appears to be due to a decrease of cyclin proteins D1 and D2.

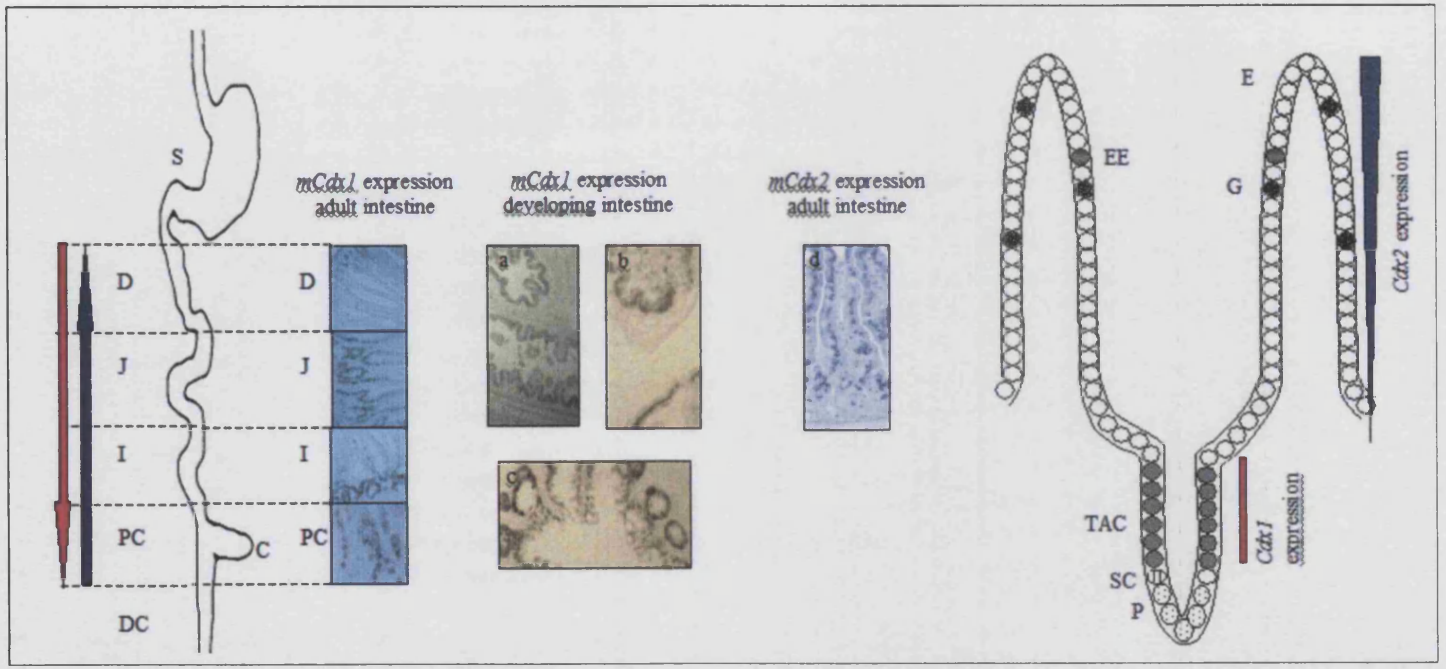


Figure 1.2. Expression of *mCdx1* and *mCdx2* in the intestinal epithelium. Stomach (S), duodenum (D), jejunum (J) ileum (I), proximal colon (PC), distal colon (DC), caecum (C). Red arrow and red bar indicate the expression of *Cdx1* in the anterior posterior axis and in the crypt villi axis. Blue arrows indicate *Cdx2* expression along both axes. Enteroendocrine cells (EE), transit amplifying cells (TAC), stem cell (SC), paneth cell (P), goblet cell (G) and epithelium (E). (a) *Cdx1* expression in small and large intestine day 16, (b) *Cdx1* expression in small and large intestine day 17, (c) *Cdx1* expression in colon day 17 and (d) *Cdx2* expression in small and large intestine. Photographs of *Cdx1* and *Cdx2* expression in the adult and in developing intestine from Subramanian et. al. (1998).

1.2.3 Genes involved in the intestinal epithelial formation and differentiation

Signaling cascades and transcription factors are required for the generation of distinct tissues and organs during development. The cascades are initiated and mediated by soluble factors like the Wnt pathway. In addition, development of tissues depends on the expression and activity of a number of transcription factors associated with these signal transduction pathways. Studies in the specification, regulation and proliferation of intestinal epithelium in the mouse have shown that several genes are involved in these complex processes. Many transcription factors, including the homeobox factors are involved in tumorigenesis in the intestinal epithelium (Clatworthy and Subramanian 2001).

Along with the expression of the *Cdx1* and *Cdx2*, other genes have been identified in maintaining the structure and function of the intestinal epithelium. The *Tcf4* factor, a member of the *TCF/LEF* family, is expressed in the proliferative region in the small intestine (Korinek *et al.* 1998). The *Cdx1* transcription factor as mentioned before, has an early expression in the proliferating epithelium; its expression is later restricted to the proliferative regions in the crypts (Duprey *et al.* 1988; Subramanian *et al.* 1998). Studies have shown that *Cdx1* stimulates cell proliferation and induces cell differentiation from the stem cells to the proliferating and transit cells. In contrast, *Cdx2* has been proved to decrease the levels of proliferating cells and enhance differentiation (Suh and G. 1996; Lorentz *et al.* 1997; Mallo *et al.* 1998; Soubeyran *et al.* 1999)

The *Fkh6* gene is expressed in the layer adjacent to the epithelium, the mesenchyme (Kaestner *et al.* 1997). *HFH11* and *HNF3β*, which belong to the forkhead family, are expressed in the crypts in the adult; however, *HFH11* is also expressed in epithelium and mesenchyme in early stages of the intestine development (Ye *et al.* 1997). And *Cdx2*, the other member of the *Caudal* family, is expressed in the villus (James and Kazenwadel 1991).

The generation of the gut involves the interaction of epithelium and mesenchyme. The basement membrane acts as a barrier between these two layers. This basement membrane is involved in cell adhesion and cytoskeletal organization. It also influences differentiation and proliferation of cells (Clatworthy and Subramanian 2001).

The Wnt signalling pathway is required for the organization of the intestinal epithelium. When the Wnt pathway is activated, β-catenin is stabilized and translocated into the nucleus. Interaction with the TCF/LEF transcription factor complex leads to the activation of genes involved in the maintenance of the gut epithelium (Brantjes *et al.* 2002). Pownall et al. (1996) and Isaacs et al. (1998) have shown that *Xcad3* is an intermediate early target of the FGF signaling pathway in *Xenopus* development, that normal expression of posterior *Hox* genes is dependent on FGF signalling, and that this regulation is probably mediated by the activation of *Xcad3*.

1.3 The Drosophila Caudal gene, expression and function

1.3.1 Anterior– posterior axis in *Drosophila*

Pattern formation and specification in the early development of *Drosophila* is initiated by a group of homeobox transcription factors expressed either maternally or zygotically in the embryo. One of these homeobox factors is *Caudal* (*cad*) that is expressed during the formation of the unfecunded egg (oogenesis) as a maternal transcript and later in development is expressed zygotically in the embryo. By the stage that the nuclei migrate to the periphery of the zygote, *cad* concentrates in the posterior end of the zygote forming an anterior posterior gradient. By stage 13, both maternal and zygotic *cad* maintain the antero-posterior gradient. Before the cellularisation stage, the maternal *cad* decreases and zygotic *cad* emerges to form first an anterior posterior gradient, that later becomes restricted to a stripe of expression located in the middle of the embryo, where the abdominal segments will be formed. This process is controlled by *hunchback* (*hb*), a gene also expressed maternally and zygotically in the embryo. The zygotic *hb* gradient regulates positively the zygotic *cad* expression in the abdominal zone of the embryo, where *cad* becomes a stripe of expression in the posterior end of *hb* gradient.

The first gradient showed by *cad* is due to the presence of *Bicoid* (*bcd*), which forms an anterior posterior gradient where the highest concentration is located in the anterior pole of the embryo, the same gradient showed by *hb*, and prevents translation of *cad*. The way in which *bcd* represses the expression of *cad* is via interaction with *cis*- elements located in the 3' untranslated region (UTR) of *cad* mRNA (Niessing *et*

al. 2002). As a result, *cad* concentrates in the posterior end of the embryo where *bcd* is not present and can not be translated. Figure 2 illustrates the distribution and gradient formed by *bcd* and *cad* during the early development of *Drosophila*. The *cad* maternal messenger is present evenly all throughout the egg (Figure 1.3a); after fertilization, *cad* mRNA locates in the posterior of the embryo (Figure 1.3b); the *bcd* maternal messenger forms an anterior gradient in the embryo (Figure 1.3c); which will create an opposite gradient to *cad*, establishing the positional identity of the anterior posterior axis en the embryo (Figure 1.3d).

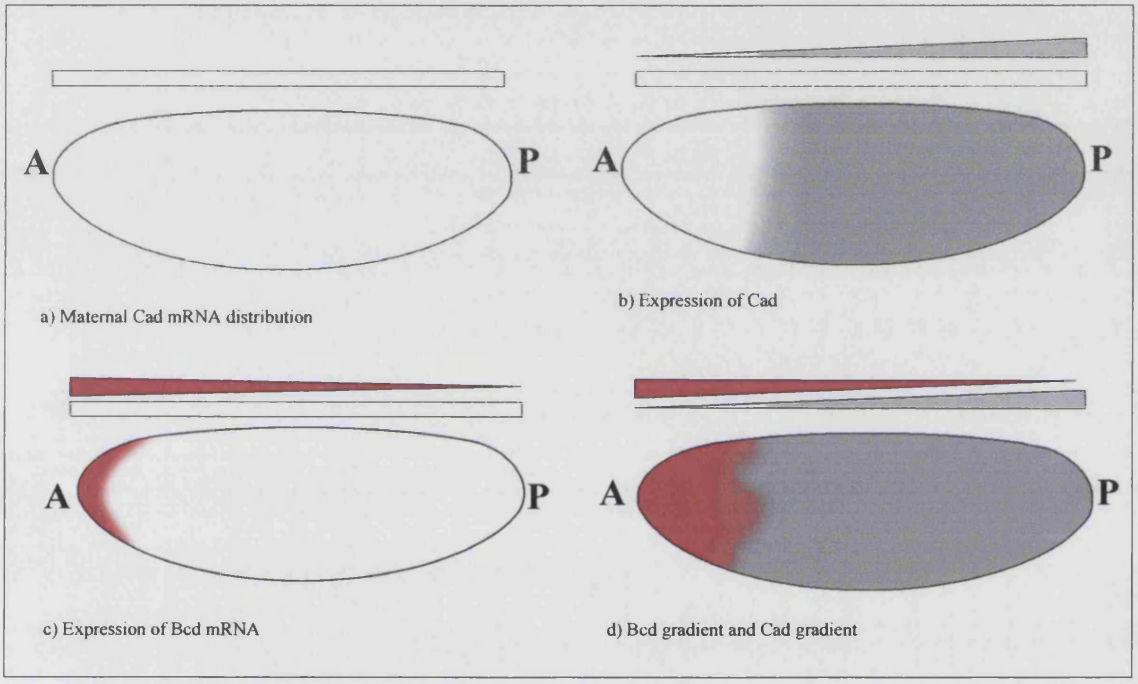


Figure 1.3. *Cad* gradient formation in the *Drosophila* developing embryo. **a)** Maternal *cad* mRNA distribution (white bar). **b)** Expression of *cad* mRNA in wild type embryo (grey bar). **c)** Expression of *bcd* mRNA (red bar) and **d)** *bcd* gradient overlapping the *cad* gradient. (red and grey bars). Adapted from Rivera-Pomar et. al. (1996).

1.3.2 *Caudal* mutations

Studies performed with *Drosophila* mutants suggest that *Cad* is also involved during gastrulation in the *Drosophila* embryo. Absence of maternal and zygotic *cad* causes defects in the development of the hindgut. When only maternal *cad* is present, there is formation of the hindgut. However, when only the zygotic *cad* is present, the

hindgut is malformed. In the absence of both, there is no formation of hindgut in the embryo. There is a redundancy in activity between the maternal and zygotic product, since the absence of one of the products can be compensated by the other (Struhl and White 1985).

The expression of other genes has been studied in *cad* mutants. Expression of *wg*, a segment polarity gene involved in the establishment of parasegment boundaries, is affected. The *fkh* gene, required for the normal development of the hindgut primordium, is also affected as well as *fog*, a gene required in the spermatogenesis process (Wu and Lengyel 1998).

1.3.3 The *Caudal* homologs: expression and function

Since the identification of the *Drosophila cad* gene, many *cad* homologs have been isolated in different species. *In situ* hybridization and immunostaining experiments have showed the distribution and expression pattern of the *Cdx* genes at different developmental stages in different organisms. These data illustrate an overlapping pattern of *Cdx* expression along the A-P axis of the embryo, from the gastrulation stage to the tail bud stage of the embryo. The expression patterns of the *Cdx* members in different stages of development in mouse, zebrafish, and *Xenopus* are described in this section and in figure 3.

Cdx1 is expressed at day 7.5 dpc. in the late primitive streak in the mouse embryo (Figure 1.4, panel A), in ectodermal and mesodermal cells, in visceral ectoderm but not expression is present in the definitive endoderm (Meyer and Gruss 1993). By 8.5 dpc, expression of *Cdx1* is present in the neural tube, somites mesoderm and limb buds (Figure 1.4, panel B). By 9.5 dpc expression is found in the dorsal region of the tail bud (Figure 1.4, panel C); at 10.5 dpc (Figure 1.4 panel D), *Cdx1* is present in the posterior tail bud, forelimb bud and anterior somites (Duprey *et al.* 1988).

Cdx2 is first expressed at preimplantation stages, in trophoectodermal cells and extraembryonic ectoderm during the implantation stage and in placenta. In later stages, at 7.5 dpc is present in the posterior primitive streak (Figure 1.4, panel A). By the day 8.5 p.c., *Cdx2* is expressed in ectoderm, mesoderm and endoderm of the embryonic tail bud; expression is also present in the neural plate, neural tube and

notochord (Figure 1.4, panel B). At 9.5 dpc, the tail bud is positive for expression (Figure 1.4, panel C) and at 12.5 dpc (Figure 1.4 panel D), expression is detected in the gut epithelium (Beck *et al.* 1995).

Expression of *Cdx4* is present at 7.5 dpc in the primitive streak (Figure 1.4, panel A); at 8.5 dpc expression is found in the hindgut endoderm (Figure 1.4, panel B) and by 9.0 to 9.5 dpc (Figure 1.4, panel C), the tail bud is positive for expression (Gamer and Wright 1993).

The presence and expression of *Cdx* has also been studied in *Danio rerio* (zebrafish). *ZfCdx* is first detected at 50% epiboly, the early stage of gastrulation in zebrafish. Expression is present in the blastoderm margin, with more restricted expression in the epiblast (Figure 1.4, panel A). At 70% epiboly, also known as the shield stage, expression is detected in the vegetal epiblast. By the end of gastrulation (90% epiboly), expression is restricted to the vegetal epiblast with a weaker expression in the hypoblast. The vegetal epiblast will constitute the future posterior part of the embryo; cells from the hypoblast will give rise to the endoderm and mesoderm tissues. During the 1 somite stage (10hrs), just after epiboly, expression is detected in ventral mesenchyme, posterior neural plate, segmental plate and the posterior tail (Figure 1.4, panel B). During mid somitogenesis at the stage of 14 somites (16 hrs), *zfCdx* is present in posterior spinal cord, notochord, ventral mesenchyme and tail bud (Figure 1.4, panel C). By the late somitogenesis stage (22hrs), expression is located in the neurectodermal cells, spinal cord, and tail bud; the mesoderm remains negative for expression. At 30hrs, also known as the pharyngula stage, expression of *zfCdx* is just detectable in the posterior spinal cord (Figure 1.4, panel D). After 48hrs, expression of *zfCdx* is undetectable in the developing embryo (Joly *et al.* 1992).

Xenopus studies have shown that *Cdx* genes are also expressed during the early development to specify the anterior posterior axis, and later in the developing gut. As in the case of the mouse, all three *Xcad* genes are expressed during the gastrulation period; *Xcad1* and *Xcad2* are expressed in the dorsal lip during mid gastrulation (stage 11); the expression of *Xcad3* at this stage is expressed in the marginal zone (Figure 1.4, panel A), although this expression becomes a broader band around the blastopore. The blastopore is a circular invagination in the ventral side of the embryo where mesoderm and endoderm converge during gastrulation. At stage 13

once the blastopore closes, expression of *Xcad1* and *Xcad2* is located in the posterior part of the embryo; *Xcad3* shows a weaker expression also in the posterior end (Figure 1.4, panel B).

During neurulation (stage17), the three *Xcad* genes are expressed in the posterior part, in the developing neural tube although the anterior expression varies for each gene at this point (Figure 1.4, panel C). Nevertheless, the anterior boundaries for each gene are visible at the 15 somite stage (24 stage), *Xcad1* extends its expression to the 5^{th} to 6^{th} somite, *Xcad2* to 7^{th}- 8^{th} somite and *Xcad3* shows its anterior expression in the 2^{nd} to 3^{rd} somite. In the posterior end, the three genes show an overlapping expression. By stage 31, expression is almost undetectable *Xcad1* shows a faint expression in the posterior spinal cord (Figure 1.4, panel D). By the 41-stage expression of *Xcad1* and *Xcad2* can be detected in the gut endoderm (Pillemer *et al.* 1998).

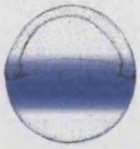
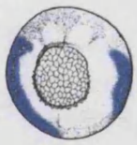
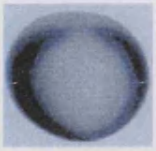
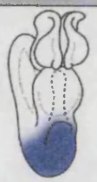
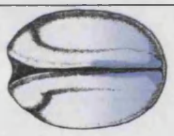
	mCdx1	*mCdx2*	*mCdx4*	*ZfCad*	*Xcad1*	*Xcad2*	*Xcad3*
Stage	 7.5 p.c.	 7.5 p.c.	 7.5 p.c.	 50% epiboly	 11	 11	 11
Expression							
Stage	 8.5 p.c.	 8.5 p.c.	 8.5 p.c.	 1 somite	 13	 13	 13
Expression	 A → P	 Tail bud section	 A → P	 A → P			
Stage	 9.5 p.c.	 9.5 p.c.	 9.0 p.c.	 14 somites	 17	 17	 17
Expression	 A → P	 Saggital section	 A → P	 A → P	 A → P	A → P	A → P
Stage	10.5 p.c.	Not Shown 12.5 p.c.	Not Shown	Segmentation /pharyngula	31	31	31
Expression	A → P	Transverse section	Not Shown	A → P	A → P	A → P	A → P

Figure 1.4. Expression of *Cdx* genes at different stages of the embryo development. Expression of the mouse *Cdx1* (*mCdx1*), *Cdx2* (*mCdx2*), *Cdx4* (*mCdx4*), zebrafish *Cad1* (*Zfcad*), Xenopus caudal 1 (*Xcad1*), Xenopus caudal (*Xcad2*) and Xenopus caudal (*Xcad3*) **A**. Represents the *Cdx* expression of *mCdx1*, *mCdx2* and *mCdx4* at 7.5 p.c., *Zfcad* at 50% epiboly, *Xcad1*, *Xcad2* and *Xcad3* at 11 stage in embryo. **B**. Shows the *Cdx* expression of *mCdx1*, *mCdx2* and *mCdx4* at 8.5 p.c., *Zfcad* at 1 somite stage, *Xcad1*, *Xcad2* and *Xcad3* at 13 stage in embryo. **C**. Shows the *Cdx* expression of *mCdx1*, *mCdx4* at 9.0 p.c., *mCdx2* at 9.5 p.c., *Zfcad* at 14 somites stage, *Xcad1*, *Xcad2* and *Xcad3* at 17 stage in embryo. **D**. *Cdx* expression of *mCdx1* at 10.5, *Zfcad* at segmentation/ pharyngula stage, *Xcad1*, *Xcad2* and *Xcad3* at 31 stage in embryo. Photographs taken from: Meyer and Gruss (1993); Beck F, et. al. (1995); Gamer and Wright (1993); James, et. al. (1994); Joly, et. al. (1992) and Pillemer, et. al. (1998).

1.4 The phenotype of Cdx1 and Cdx2 knockout mice

1.4.1 The *Cdx1* $^{-/-}$, effects on axial identities

In accordance with the expression of the *Cdx* genes during the establishment of the anterior posterior axis and their regulation of *Hox* genes, the *Cdx1* homozygous mutant shows skeletal abnormalities; the 1-7 cervical vertebrae and the first thoracic vertebrae suffer an anterior transformation in the mutant mice, the *Cdx1*$^{-/-}$ also shows that the *Hox* genes undergo posterior shifts in their expression domains. This phenotype is similar to those shown by the disruption of the *Hox* genes. The *Hoxd-3* mutant mice show fusion of the basioccipital bone with the first vertebrae. This is, the anterior arch of the atlas transformed into the bassioccipital bone. The axis (the second vertebrae) shows atlas characteristics. Other vertebrae are also affected either because some of the vertebrae parts are deleted or because they are transformed to an anterior part (Condie and Capecchi 1993). Furthermore, the homeotic anterior transformations described for the *Hox* mutants are located in specific region whereas in *Cdx1*, the transformations occur in a wider area. These findings confirm that *Cdx1* is necessary to delimit the expression boundaries of the *Hox* genes in the primitive streak (Subramanian *et al.* 1995).

1.4.2 The *Cdx2* $^{-/+}$ mice, intestinal tumour formation

The *Cdx2* null mutant mouse was generated by homologous recombination (Chawengsaksophak *et al.* 1997); the homozygote null mutant dies during the implantation stage between 3.5 and 5.5 dpc. However, the *Cdx2* heterozygote mutants are viable and display a variable phenotype. At early stages, the mutant mice exhibit skeletal abnormalities; homeotic transformations are observed from the sixth cervical vertebrae to the eight thoracic vertebrae. The mutant mice are also characterized by a short tail; these axial abnormalities are linked to alterations in the *Hox* genes expression boundaries. In the first three months of development, a high percentage of the heterozygote mice develop multiple polyp-like lesions in the gut; the highest frequency of polyps are located in the proximal colon, in the area where *Cdx2* is usually expressed. No polyp-like lesions are observed in regions where *Cdx2* is not expressed. These polyps consist of normal gastric mucosa with villi and Paneth cells of intestinal tissue. The neoplastic cells present in the transformed intestinal epithelium do not express *Cdx2*.

1.4.3 Overlapping function of the *Cdx1* $^{-/-}$ *Cdx2* $^{-/+}$ genes

The *Cdx2*$^{+/-}$/ *Cdx1*$^{-/-}$ double mutants also show skeletal abnormalities in the vertebral column. In this case, the homeotic transformations are more severe than the single mutant phenotypes and the posterior shift in the expression of the *Hox* genes is more affected, suggesting that the *Cdx* genes have an overlapping role in the anterior posterior patterning via regulation of *Hox* gene expression (van den Akker *et al.* 2002).

As mentioned before, the second *Cdx* expression phase is in the developing and adult intestinal epithelium. In *Cdx2+/-*, the intestine acquires a forestomach epithelium in the midgut; this ectopic tissue is able to differentiate into all the tissues of the stomach but is unable to produce small intestinal tissue, indicating that *Cdx2* is required for the maintenance of the intestinal epithelium (Beck *et al.* 2003).

Studies have related *Cdx1* and *Cdx2* to colorectal tumors. *Cdx2* decreases with the advance of the tumor and *Cdx1* is upregulated in the early stages of the tumor development. Further studies have argued that *Cdx1* has oncogenic properties. Firstly

ectopic expression of *Cdx1* is present in stomach, liver and oesophagus adenomas; furthermore, *Cdx1* is upregulated by the oncogenic *Ras* and the *Wnt/β- catenin* signaling pathways. However *Cdx1* is down regulated in colorectal tumors (Mallo *et al.* 1998; Lorentz *et al.* 1999).

Along with the *Cdx* genes, several genes are also expressed in the gut with characteristic expression patterns in the normal colon and during tumor progression. *p53*, a gene expressed in the base of the crypts, has been shown to downregulate *Cdx1* in IEC-6 cells. *p53* expression is lost during tumor progression, which might allow an increase in *Cdx1* expression in the colon. Assays performed in IEC-6 and SW480 cells show that *Cdx1* also stimulates the *Bcl-2* gene, another factor expressed in the stem cell reservoir of the crypt (Moucadel *et al.* 2002). The same study shows that *p21*, a gene expressed at the top of the crypt and involved in cell arrest and differentiation, is downregulated by *Cdx1*, suggesting that *Cdx1* might prevent the differentiation process in the intestine.

Sox9, a gene that is expressed at the bottom of the crypt in small intestine and colon and depends on the *Wnt/β- catenin* pathway, is able to repress *Cdx2* expression in colon carcinoma cells (Blache *et al.* 2004).

1.5 Role of Cdx in haematopoietic differentiation

1.5.1 Role of *Cdx4 in* specifying blood progenitors in zebrafish

Apart from their role in the A-P axis formation and maintenance of the intestinal epithelium, the *Caudal* family, mainly *Cdx2* and *Cdx4*, have been implicated in haematopoietic lineages. Two recent studies have identified *Cdx4* as a regulator of the haematopoietic stem cells via regulation of the *Hox* genes (Davidson *et al.* 2003), and *Cdx2* as a precursor of leukemia (Rawat *et al.* 2004).

One of these studies (Davidson *et al.* 2003) states that *Cdx4* is involved in inducing blood formation in the kgg^{tv205} mutant in zebra fish. The *kgg* mutant, who carries an autosomal recessive mutation, was first identified to produce tail defects in the zebrafish embryo. Anaemia is also present from the first day of development; these mutants survive until the 10^{th} day of development.

Expression of *scl, gata1,* and *runx1* genes in the *Kgg* mutant was shown to be altered, producing haematopoietic abnormalities. Genes involved in the formation of kidney, such as *Pax2.1* and *Cxcr4b*, showed shortened expression in these mutants. Further analysis showed that expression of *wt1* is extended from somite one to somite six, suggesting an expansion of anterior kidney fates. Structures like head, notochord and somites appear normal although the embryo is shorter in length (Davidson *et al.* 2003).

Sequence analysis showed that the *Cdx4* gene in the kgg^{tvl240} mutants contains a nucleotide transversion (T to A), which produces a change in the amino acid sequence (F170L). This mutation disrupts the protein binding to the *Cdx4*-binding site. *Cdx4* expression is not present in the nascent blood islands but its expression partially overlaps with the *Scl* in mesodermal cells. Overexpression of *Cdx4* showed ectopic expression of *Scl, Gata1* and *fli* genes near to the endogenous blood precursors at 5- 12 somite stage. Injection of *Cdx4* mRNA partially rescued the population of *Gata1* and *Scl* expressing cells in the *Kgg* mutant (Davidson *et al.* 2003).

Expression of *Hoxb4, Hoxb5a, Hoxb6b, Hoxb7a, Hoxb8a, Hoxb8b* and H*oxa9a* showed an altered expression pattern in the kgg^{tv205} mutants. The stripe of haematopoietic vascular precursors is affected by changes in the anterior posterior patterning. There is loss of $Gata1^{+}$ haematopoietic cells from the posterior stripe. Expression of *Hoxb6b*, *Hoxb7a* and *Hoxa9a* was also reduced. Over-expression of *Cdx4* was able to rescue *Hoxb6b*, *Hoxb7a* and *Hoxa9a* in *Cdx4* morphants (Davidson *et al.* 2003).

To investigate if *Cdx4* enhances the self-renewal or proliferation in stem cells, retroviral expression of *Cdx4* was assayed in the mutant. *Cdx4* induced multilineage haematopoietic progenitors, colony forming unit granulocyte erythroid/macrophage/megakarocytes colonies. *Cdx4* was also able to induce expression of *HoxB3*, *HoxB8* and *HoxA9*, genes that have been implicated in haematopoietic stem cell and immature expansion. These experiments indicate that disruption of *Cdx4* expression causes perturbation of Hox expression, causing a reduction of erythroid cells (Davidson *et al.* 2003).

1.5.2 *Cdx2* and its relevance in acute myeloid leukaemia (AML)

A study performed by Rawat et. al. (2004) relates the ectopic expression of *Cdx2* to the development of acute myeloid leukemia using a mouse model. The acute myeloid leukemia is produced by a gene fusion mechanism. The oncogenic potential of fusion genes is characteristic of the *ETV6* gene.

There are mainly two different types of translocation related to the ETV6 gene, 1) fusion to phosphotyrosine kinases and 2) fusion to transcription factors. The transcription factors involved in these translocations are *AML1* and *Cdx2*. In the *ETV6-AML1* fusion, no bone marrow transformation has been seen. In the *ETV6-Cdx2* fusion gene, ectopic expression of *Cdx2* induces AML. However, *ETV6- Cdx2* was unable to produce leukemia.

Irradiated mice transplanted with bone marrow cells expressing *Cdx2*, became moribund after post transplantation. These mice showed hematopoietic abnormalities and developed AML with formation of blast colonies in bone marrow, peripherial blood and spleen. Wild type mice injected with bone marrow cells of diseased *Cdx2* animals died within 24hrs after injection (Rawat *et al.* 2004). Conversely, irradiated mice transplanted with bone marrow cells expressing *ETV6-Cdx2* fusion, showed an increment in their hematopoietic cell populations. None of the animals suffered from anemia or blast colony formation in peripheral blood (Rawat *et al.* 2004).

This study also showed that the N-terminal and the homeodomain of the Cdx2 protein have to be intact in order to produce leukemia. Mutations produced in the PBX1 binding motif (W167A- *Cdx2*) of *Cdx2* or the simple expression of *Cdx2*, showed blast formation in bone marrow cells. Mice transplanted with these cells developed leukemia. Cells expressing the *ETV6-Cdx2* fusion, or an inactive form where the N-terminal of the *Cdx2 (ΔN-Cdx2)* was removed or a form with a mutated *Cdx2* homeodomain (*N515-Cdx2*), did not form blast cell populations (Rawat *et al.* 2004).

Analysis of the *Cdx2* and *ETV6-Cdx2* transplanted mice showed that *Cdx2* did not increase expression of the leukemogenic homeobox genes *HoxA9* and *Meis1*. Constitutive expression of *Cdx2* is highly leukenogenic when it is ectopically

expressed in hematopoietic progenitor cells. Ectopic expression of *Cdx2* might activate genes that are not expressed in blood development. The mechanism of transcriptional activation of *ETV6- Cdx2* is still to be identified; *ETV6* promotor/enhancer might be involved in the activation of *Cdx2* in AML.

1.6 Regulation of the mouse Cdx1

1.6.1 Regulation of the mouse *Cdx1* by the Wnt signalling pathway

Due to its role in proliferation of the cell population in the intestinal epithelium, and its link with the development of tumors in the colon, *Cdx1* has been a gene of interest in the study and understanding of structure and maintenance of the intestine as well as in the study of the beginning of intestinal cancer. Recent studies have aimed to understand the regulation of the *Cdx1* gene by attempting to identify the regulatory elements involved in its regulation. Most of the work has been done using the murine *Cdx1* non-coding regions to test for enhancer regulatory regions either in cell lines or producing transgenic mice. This section describes the work and the understanding that we have to date in the regulation of this gene.

The importance of the Wnt pathway in diverse developmental processes has been demonstrated in a variety of organisms. Lickert et al. (2000) showed that the mouse *Cdx1* is a direct target of the *Wnt/β- catenin* signalling pathway. The first findings of the regulation of *Cdx1* by Wnt were achieved using ES cells co-cultured on 3T3 cells expressing *Wnt. Wnt1, Wnt3a, Wnt7a* stimulated the expression of *Cdx1*; however the strongest induction was given by *Wnt4.*

Deletion promoter analysis showed that the –3.6Kb non-coding sequence of the *Cdx1* gene was able to drive the expression of the gene in HEK293 cells. Four TBE (Tcf binding elements) were found to be contained in this 3.6Kb sequence, two of these sites located in the first –350bp.

Mutations of these four TBE motifs and combination of these mutations showed that the elements situated in the first 350 bp were relevant for the expression of the reporter. Electromobility shift assays proved that the β-catenin/*Lef1* complex was able to bind to the two TBE sites present in the *Cdx1* upstream region.

Expression of *Cdx1* was observed in the co-cultures of small intestinal endoderm from rat embryos and 3T3 cells expressing *Wnt*, suggesting that *Wnt1* can induce the expression of *Cdx1* in embryonic endoderm.

Analysis of *Tcf4-/-* embryos, which lack proliferative cells in the crypt region of the small intestine, showed a reduced staining with the anti- *Cdx1* antibody at 15.5 dpc; by 17.0 dpc, no *Cdx1* expression was visible in the crypt region of the small intestine, although *Cdx1* expression was positive in the colon epithelium of the *Tcf4-/-* embryos.

1.6.2 Regulation of *mCdx1* by TBE and RARE sites

Once the regulation of *Cdx1* was established via the Wnt/β-catenin pathway, the candidate regulators for the *Cdx1* transcription were the *Tcf* factors; reporter constructs carrying different lengths of the *Cdx1* upstream region were used to create transgenic mice. Analyses of the regions that showed expression of the reporter were analyzed for transcription factor binding sites; the TBE and retinoic acid response elements (RARE) were identified as responsible of the regulation of *Cdx1*.

Transgenic mice carrying the –3.6Kb non-coding region of *Cdx1* were found to express the transgene in mesodermal and ectodermal cells of the primitive streak at 7.5 dpc. By 8.7 dpc, expression in mesoderm and ectoderm was detected. Neural tube, somites and tail bud mesoderm also showed expression of the reporter. At 9.5 dpc, expression in the anterior region was present in the migrating neural crest cells, and dorsal root ganglia. In the posterior region, expression of the transgene was in the neural plate and tail bud mesoderm. No expression was seen in the hindgut endoderm. By 14.5 dpc, no reporter expression was detected in the embryonic intestine (Lickert and Kemler 2002).

Furthermore, the 0.7Kb upstream of the *Cdx1* gene was able to drive the expression of the transgene in a similar fashion to the one seen in the –3.6Kb construct at 7.5 dpc. By 8.5 dpc, expression was present in the neural tube, the border of the hindbrain, somites and paraxial mesoderm. By 9.5 dpc, expression was detectable in neural tube, somites and paraxial mesoderm; however, no expression was present in the hindgut ectoderm. Mutation of the TBE and RARE sites

demonstrated that when the RARE site was mutated and the TBE sites were left intact, expression was observed in the primitive streak with enhanced expression in the base of the allantoids at 7.5 dpc. By 7.75 dpc, expression was seen in the ectoderm and posterior region of the embryo. By 8.5 dpc, expression was just present in the base of the allantoids, low expression was observed in the ectoderm and neural plate, and no expression was seen in the mesoderm. When the 2TBE and the RARE were mutated, no expression was present in the embryo (7.5 to 8.5 dpc).

The –0.7Kb promoter element seems to contain all the necessary cis-regulatory elements for the expression and regulation of the mouse *Cdx1* in the early stages of development. The TBE sites are important for the *Cdx1* expression at 7.5 dpc and the RARE site is essential for initiation of expression in the primitive streak mesoderm and for the maintenance of expression in the primitive streak ectoderm.

1.6.3 The *CdxA* regulation, the conserved TBE and RARE elements

Gaunt et al. (2003) used the chick *Cdx1* (*CdxA*) to produce transgenic mouse embryos. A reporter construct containing 1.4Kb upstream region, 1st exon and 2.1Kb of the first intron was able to drive the expression of the *LacZ* gene from the adjacent region of the neural tube to the first somites at 8.5 dpc. By 8.7 dpc, strong expression is seen in the neural tube, whereas a weak staining is detected in the paraxial mesoderm. However, an extra construct containing 6Kb upstream region, part of the first exon and the 2nd intron of the β-globin, showed no reporter expression for the eight transgenic lines produced.

An analysis of the *CdxA* sequence showed the presence of one RARE and a *Tcf/β-catenin* sites present in the 1st intron of the gene. Mutation of sites proved that when the RARE site is destroyed expression is seen in the neural tube at 8.7 dpc; however the boundary of expression is posterior to the somite 5. When the *Tcf/β-catenin* is mutated, expression is seen in few cells in the posterior neuroectoderm, and the boundary of expression is until the somite 13. No expression is seen in the mesoderm at 8.7dpc.

To establish if the expression pattern showed by the *CdxA* element is similar to the one produced by the *Cdx1*, a reporter construct containing 5.7Kb mouse DNA

(3.3Kb upstream region, 1st exon, and 1.8Kb of the 1st exon approx.) was used to produce transgenic mice. Expression in the mesoderm and neuroectoderm was observed at 8 dpc. By 8.7 dpc, a gradual expression was seen in these tissues; expression was also present in the anterior boundaries at the level of the first somite in the neuroectoderm and in the somitic mesoderm at the level of the fifth somite. Neural tube and lateral plate mesoderm were positive for the expression. At this stage, weak expression was seen at the level of the fith somite, due to a time- gradient expression.

In contrast with the work presented by Lickert and Kelmer (2002), Gaunt and collaborators showed that the cis regulatory regions involved in the regulation of *CdxA* are contained in the first intronic region instead of the upstream region of the *Cdx1*. However, the RARE and *Tcf/β-catenin* motifs seem to play a crucial role in the early expression and regulation of the gene.

1.6.4 Regulation of the human *Cdx1*

More recent studies have shown that the upstream region of the human *CDX1* contains the necessary elements to regulate the expression of the gene (Rankin *et al.* 2004). Reporter constructs carrying the *LacZ* gene were used to generate transgenic mice. A –327 to + 68 bp element drove reporter expression in the base of the crypts in the small intestine and a scattered expression in the colon. However, different founder lines showed different patterns in expression or no detectable expression at all. These indicate that this heterogeneous expression is due to the site insertion of the transgene.

The region contained between –5667 and +68bp of the *hCDX1* gene was able to drive the expression in the serosa of the small intestine and colon, submucosa muscle layer, stomach, squamous cells of the kidney and granular cells of the cerebellum. Furthermore, the –15601 to +68bp region showed strong expression in intestinal epithelial cells, ileum and colon; no expression was detected in any other tissue. These suggest that the expression of this construct is insertion independent. By 10.5dpc, this construct was expressed in the somites, neural tube, limb bud, tail bud, caudal region and embryonic intestine. By the 12.5 dpc a gradual decrease in expression was observed in the ectoderm and mesoderm. By E13.5, strong expression was seen in the intestinal epithelium and by 14.5 dpc strong expression in the midgut

(jejunum and ileum), foregut and hindgut showed a weak expression and the stomach was negative for the staining.

In conclusion, the –5667 and +68bp region seems to be responsible of the *hCDX1* expression during the early (mesoderm and ectoderm) and later (endoderm) stages of development. In addition, DNase I hypersensitive assays revealed two putative intestinal enhancer positioned at –5.8 and -6.8Kb upstream of the gene.

1.6.5 Regulation of *mCdx1* by Retinoic Acid

Studies have shown that hormone receptors mediate the effects of hormones on several genes. Retinoic Acid (RA) acting through the RA receptors (RARs) and Wnts have been implicated in vertebral patterning (Kessel and Gruss 1991). It is well established that RA and RAR play important roles in differentiation, growth, and homeostasis of epithelial cells in various tissues (Kastner *et al.* 1995; Mangelsdorf *et al.* 1995). RAR belongs to a large steroid/non steroid nuclear hormone receptor superfamily that consists of three receptor isotypes α, β, γ, each encoded by distinct genes. RAR forms a heterodimer with Retinoic X Receptors (RXRs) to bind to the RARE on the target genes. On its own, RAR DNA binding activity is weak, but it is enhanced by RAR/RXR dimerization (Yang *et al.* 1991). RAR consists of a DNA-binding domain that contains various functional domains. RAR interacts with other transcription factors such as CBP, p300 and nuclear receptor activators, which participate in DNA remodelling.

RA has been implicated in the regulation of *Cdx1*. Embryos treated with retinoic acid showed an increased of *Cdx1* expression in the posterior regions (Houle *et al.* 2000). The same up regulation is seen in embryo cultures and embryonal carcinoma cells. In embryos treated with RA, *Cdx1* expression is induced in the complete region of the primitive streak at 7.5 dpc. At 9.5 dpc, treated embryos showed an increase of expression in the fore limb buds, mesenchyme and presumptive dermamyotome.

The RARα1 $^{-/-}$ mice when treated with retinoic acid, showed an induction of the *Cdx1* transcript at 8.5 dpc; however, in the treated RARα1/γ mutant, that induction in expression was compromised. Analysis of the -2Kb *mCdx1* fragment

containing the internal ATG and the 5'UTR of the gene was able to drive the expression of the reporter in F9 carcinoma cells. The same activity was seen when the ATG and 5'UTR were removed. Stable cell lines carrying the remaining region were capable to respond to small doses of retinoic acid, which resembles the response to RA *in vivo*.

Analysis of this sequence showed the presence of a RARE element between –694 and -185 relative to the transcription start site. By EMSAs, this element was shown to be target for RAR and RXR factors. Mutation of the RARE site reduced the expression of the reporter in F9 cells. In agreement with the role of RA regulating *Cdx1* and the data obtained with the RARα1/γ mutant, it is suggested that the role of retinoic acid in the control of *Cdx1* expression is as an initiation factor, and the presence of other factor (s) may be the target in the maintenance of gene expression in later stages.

1.6.6 RARE regulates the expression of *mCdx1* in early development

Further studies by Houle et al. (2000; 2003) gave more insights in to the interaction of RA and the *Cdx1* regulatory region. Reporter assays using F9 cells proved that RA induced the expression of the $RARE^{Wt}$ whereas no expression was induced in the $RARE^{Mut}$.

The RARE in the 2Kb upstream of the *Cdx1* was inactivated using a floxed neomycin selection cassette and the mutated RARE. The mice carrying the $RARE^{Mut}$ showed a reduced presence of the *Cdx1* transcript. A weak expression was evident in the early somites stages with almost null expression in later stages. Although the *Cdx1* expression is reduced in the neuroectoderm at 8.0 dpc, no effect was seen in the rostral boundary expression, suggesting that the RARE site is necessary for the early and later stages of expression (Houle *et al.* 2003).

To investigate if the $RARE^{Mut}$ responds to RA *in vivo*, pregnant mice were treated with RA. The wild type embryo exhibited an increase of the transcript in the primitive streak at 7.0 dpc whereas the RARE mutants showed the same level of expression at the same stage. From 7.5 to 8.5 dpc, $RARE^{Wt}$ and $RARE^{Mut}$ reflected

the same level of induction suggesting that the regulation of *Cdx1* is via another RA pathway, where RARE is not involved in the regulation of *Cdx1* (Houle *et al.* 2003).

1.6.7 Regulation of genes by RA and RAR

Retinoic acid, a derivative form of vitamin A, has been implicated in the regulation of the *Hox* genes. Studies have shown that RA can induce normal sequence of expression of the *Hox* genes (Boncinelli *et al.* 1991). Further studies demonstrated that a dose of exogenous RA provokes anteriorization of the *Hox* genes expression in mouse embryo that resembles posterior homeotic transformations (Kessel and Gruss 1991).

The mode of RA action is by binding to its receptors RARα, RARβ and RARγ; these nuclear receptors regulate the expression of genes by forming heterodimers with the retinoic X receptors. These heterodimers bind to the RAREs present in the promoter of genes which are inducible by RA (Mangelsdorf *et al.* 1995; Chambon 1996). The RARE motif is a repeat sequence of PuG(G/T)TCA however, this sequence has revealed to be highly polymorphic.

Studies have described the RAREs as a common element in the activation of *Hox* promoters, demonstrating that these genes are direct targets of RA. Expression of *Hox* genes with a RARE element has been described in the hindbrain, neuroectoderm and paraxial mesoderm and also in mesoderm like in the case of the *Hoxd4* gene (Zhang *et al.* 1997).

RA is not only involved at the level of *Hox* gene expression, its activity sets out earlier than the activation of *Hox* genes. The *Cdx1*$^{-/-}$ mouse displays homeotic transformations of the axial skeleton, which are very close to the RARα/γ- null phenotype and a RARE regulates expression of *Cdx1* in tissue culture (Lohnes *et al.* 1994; Subramanian *et al.* 1995; Houle *et al.* 2000; Allan *et al.* 2001). This work has proved that RA is a regulator of *Cdx1* in the early development.

1.7 The Adenomatous Polyposis Coll (APC) gene

1.7.1 Expression and regulation of the *Apc* gene

Alteration of cell proliferation and differentiation is thought to be a major event in the initiation and progression of cancer. Colorectal cancer is a multi-step process linked to the inactivation or loss of expression of tumor suppressor genes. In general, an intestinal cell needs to comply with two essential requirements to develop into a cancer: it must acquire a) a selective advantage to allow for the initial clonal expansion, and b) genetic stability to allow for multiple hits at other genes responsible for malignant transformation (Fodde 2002).

Colorectal tumors are known to arise through a gradual series of histological changes, the so-called adenoma carcinoma sequence. Each change is accompanied by a genetic alteration in a specific oncogene or tumor suppressor gene. Loss of Apc function triggers this chain of molecular and histological changes. Extracolonic manifestations, including gastric and duodenal polyps, osteomas, desmoids, epidermoid cysts and retinal lesions, are commonly observed in patients with familial adenomatous polyps (Santoro and Groden 1997).

The Adenomatous polyposis coli (APC) protein is one of the key players in the (*Wnt*) pathway and regulation of *β- catenin.* The *Wnt* signaling pathway is a very well conserved process across species that regulates proliferation and differentiation of cells. The role of APC in this signaling pathway is to regulate the β-catenin levels in the cytoplasm. In the absence of Wnt molecules, APC forms a complex with Axin, PP2A and GSK3β; the role of Axin, or its homologue Conductin, is to form the scaffold in this multiprotein complex, GSK3β phosphorylates β-catenin in its N-terminal region leading to the ubiquitination of β-catenin by the β-transducin repeat-containing protein (β-TRCP) and subsequent degradation by the proteosome pathway (Polakis 2000).

In the presence of Wnt signal, Frizzled, a transmembrane protein, activates Dishevelled (Dsh). Although it is not known how Dsh is activated and its secondary steps, it interacts with casein kinases like CK1, CK2 and GBP/Frat1, inhibitors of β-catenin. Dsh binds to Axin resulting in the disruption of the multimeric complex and accumulation of β- catenin in the cytoplasm (Figure 1.5, panel B). β- catenin

translocates into the nucleus to bind to the *TCF/Lef* complex. *TCF* and *Lef* transcription factors are members of the lymphoid high mobility group (HMG). This results in the activation of genes involved in cell cycle, proliferation and differentiation of cells such as *c-myc*, *Cyclin D1*, *TCF1*, *conexin 43*, *metalloproteinase matrylsin* and *Cdx1* (van der Heyden *et al.* 1997; He *et al.* 1998; Crawford *et al.* 1999; Shtutman *et al.* 1999). The *Wnt* pathway also regulates the expression of the Eph/ Ephrin surface molecules, which specify the position of each epithelial cell along the crypt villus axis (Batlle *et al.* 2002).

In the absence of Wnt signal and β-catenin, the *TCF/Lef* complex is inhibited by transcriptional co- repressors like the TLE/Groucho family or C-terminal binding protein CtBP (Brannon et al. 1999). Disruption of the Wnt signaling pathway can lead to an increase in proliferation and loss of differentiation (Figure 1.5, panel A). Thus, deregulation of APC produces inappropriate transcription, cell adhesion and disruption of the pathway. Mutations in the components of the Wnt pathway, such as the Apc or β-catenin, result in a continous translocation of β-catenin in to the nucleus, which in turn produces a constitutive expression of genes commonly found in colon cancers (Morin *et al.* 1997).

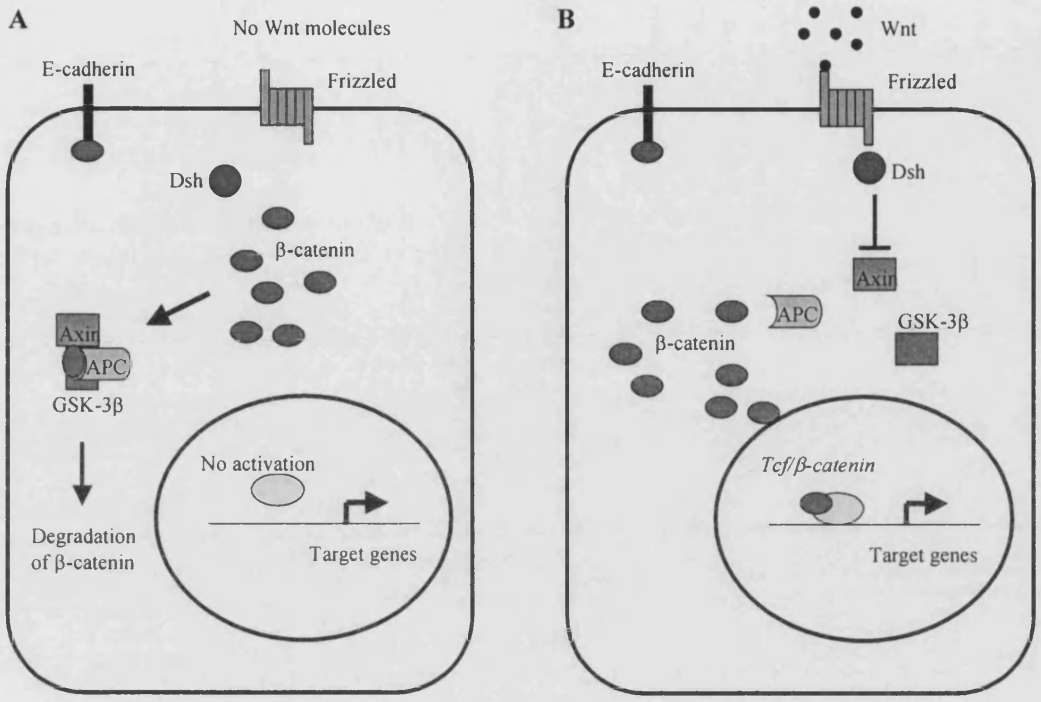


Figure. 1.5. The Wnt-signaling pathway. A. In the absence of a Wnt signal, β-catenin is phosphorylated by the GSK3β/APC/Axin complex resulting in degradation of β-catenin.Transcription of genes is blocked due to the absence of Tcf/β- catenin complex in the nucleus (A). In the presence of a Wnt signal, Dishevelled (Dsh) inactivates Axin and GSK3β. The Axin/APC/ GSK3β is inhibited and β- catenin translocates into the nucleus to for a complex with *Tcf* factors to activate genes involved in proliferation and differentiation (B). Figure adapted from (Willert *et al.* 1997).

1.7.2 Role of APC in colon cancer

Inactivation of APC or direct inactivation of β- catenin by mutations in its APC binding site provoke accumulation of β- catenin in the cytoplasm. Studies have shown that introduction of wild type APC into cells with mutated APC protein reduces the pool of β- catenin (Korinek *et al.* 1997; He *et al.* 1998).

Colon cancer is the product of mutations that cause a continuous proliferation signal in the cells. Mutations in APC have been linked to Familial Adenomatous Polyposis patients (FAP). FAP is an autosomal dominant condition, which produces adenomatous polyps in the rectum and colon and in some cases, the polyps develop into tumors. The colonic cancer epithelial cells from FAP patients express only truncated forms of APC, lacking the central region of the protein, essential for the downregulation of β-catenin. Recently, mutations in Axin1 or Conductin, components of the destruction complex have been found in hepatocellular carcinomas and in colorectal cancer (Hart *et al.* 1998).

Outside of the Wnt pathway, APC has a role in cell migration and intercellular adhesion. Studies performed in mice carrying a mutated APC version of the protein showed that enterocytes accumulate in the crypt villus boundary. These cells contained an increased level of β- catenin, indicating that normal APC locates in this boundary allowing the cells to migrate from the crypt to the villus (Wong *et al.* 1996).

1.7.3 Role of APC in cell migration

One of the main characteristics of tumour tissues is the loss of tissue structure and architecture; cell migration and cell adhesion are affected in these type of cells. APC has been found to regulate cell migration by regulating the microtubule function. Microtubules are cytoskeletal polymers of α/β tubulin subunits, which are involved in cell migration. They serve as a track for motor proteins such as dyneins and kinesins, direct vesicle transport, locate organelles in their position and locate the microtubules themselves. APC binds to the microtubules that do not belong to the actin network and localizes at the edges of the cells (Smith *et al.* 1993; Neufeld and White 1997). Overexpression of the APC C- terminus shows its microtubule binding capacities; *in vitro* studies showed that the carboxy-terminal region of the APC protein binds to cytoplasmic microtubules, this region is usually deleted in cancers (Munemitsu *et al.* 1994).

Mahmoud *et al.* (1997) demonstrated that APC contributes to the migration of intestinal cells in the crypt-villus axis. A truncated APC protein affects the cell migration in the Min/+ mouse; no enterocyte migration, proliferation, apoptosis, or β-catenin levels are affected in these mice. Overexpression of APC results in an altered

migration of intestinal cells. APC is also able to bind to the Rac specific guanidine nucleotide exchange factor (Asef), a small G protein involved in cell migration via interaction with actin cytoskeleton (Kawasaki *et al.* 2000).

1.8 Use of the zebrafish in developmental biology

Teleosts, such as the zebrafish (*Danio rerio*), the pufferfish (*Fugu rubripes*) and the medakafish or medaka (*Oryzias latipes*), are increasingly being used as vertebrate model systems in developmental biology and comparative and functional genomics. There are two main reasons why the teleosts are used as vertebrate models in these fields: firstly, their relative ease of application in genetics and secondly, their relatively small genome size. The compact genome of the zebrafish has facilitated molecular studies of human genes because most human genes that originated early in vertebrate evolution exist in the teleost genomes. In the zebrafish in particular, large-scale mutagenesis has been achieved, which has made it highly attractive as a vertebrate model for developmental genetics (Driever *et al.* 1996; Haffter *et al.* 1996).

A major advantage of the zebrafish is that it is an optically clear embryo, which allows observation of temporal and spatial changes in gene expression during the development of the living embryo. Further advantages are the easy accessibility of embryos and the number of embryos that can be collected and manipulated in a relatively short period of time.

Many zebrafish promoters have been isolated and characterized using the green fluorescence protein (GFP) or the red fluorescence protein (RFP); some of these promoters are specific for the gut, like the intestinal type fatty acid- binding protein (I-FABP). Reporter analyses revealed that the proximal 192-bp region of the I-FABP promoter is sufficient to direct intestine-specific expression during zebrafish development. Transgenic lines have been made to use the I- FABP gene as a marker for intestinal development in the zebrafish (Her *et al.* 2004).

Similar studies have used this transient-transgenic approach to screen conserved regulatory elements for enhancer activity in the zebrafish embryo (Woolfe *et al.* 2005). Whichever is the final output, whether in the use of developmental

biology or functional genomics, the zebrafish has proved to be a very quick and useful model for *in vivo* developmental studies.

1.9 Use of the Fugu in comparative genomics

The Japanese pufferfish genome has been used for a variety of studies in the field of comparative genomics. Firstly, the *Fugu* genome was used to address the location and structure of genes in human and higher organisms due to their highly similar repertoire of genes. In particular, the *Fugu* genome has been useful for addressing conserved non-coding regions that harbor regulatory elements in their sequence. The large evolutionary distance between *Fugu* and human offers a filter for the searching of these conserved non-coding elements. Such analyses would be impossible to do between human and rodent due to the high evolutionary conservation between their genomes.

Fugu has a small genome (390Mb), 8 times smaller than mammalian genomes. Due to its small genome size, *Fugu* has a very high gene density, small introns and low percentage of repetitive sequences (Brenner *et al.* 1993).

The conserved linkage between *Fugu* and human genes has been an advantage for positional cloning of human genes, finding new coding sequences and locating the intron/exon structures of the genes. Thus, essential regulatory elements may also be conserved, although these regulatory elements may be located much closer to the genes they regulate than the ones in mammalian genomes due to the small genome size in *Fugu* (Aparicio *et al.* 1995; Elgar *et al.* 1996; Venkatesh *et al.* 1997; Gellner and Brenner 1999).

Several studies made use of this comparative tool to look for regulatory elements of developmental genes. The homeobox gene *Hoxb-1*, expressed during development in the spinal cord and hindbrain, was found to contain a conserved enhancer that directs the expression of the reporter during the early development. Similar results were found with *Hoxb-4* where three conserved non-coding regions can drive the expression of the transgene in mouse embryos. Two important points are to be considered from this work; that the conserved non- coding regions tested were tissue specific, and that some of the conserved elements used to prepare the

transgenics were *Fugu* DNA, which was able to replicate to some extent the expression of the homologous gene in the mouse embryo (Marshall *et al.* 1994; Aparicio *et al.* 1995).

Other examples have shown that the conserved elements are distributed not only upstream of the gene they regulate. The *Wnt-1* gene, that is expressed during development in the midbrain and anterior region of the hindbrain, possesses a 110 bp conserved region located in the 3' end of the *Fugu Wnt-1*, sufficient to drive the expression in the hindbrain (Rowitch *et al.* 1998).

The *Otx2* gene is essential for rostral head formation during development; a comparative and functional analysis of *Fugu Otx2* revealed seven conserved elements located over 60 Kb in the *Fugu Otx2* loci; the conserved elements were found to be independent in the regulation of expression when tested in zebrafish and mouse embryos (Kimura-Yoshida *et al.* 2004). These studies illustrate the use of comparative genomics, in specific the use of *Fugu*, in the searching of potential regulatory regions across species.

1.10 Objectives

The aims of this study are first; to understand how the *mCdx1* homeobox factor is regulated during development. Specifically, the objective was to locate the regulatory regions responsible for the early expression of the *mCdx1* and the elements involved in the late expression of the gene in the intestinal epithelium. Despite other studies on the regulation of Cdx1 factor, the complete understanding of its regulation is still unknown.

A second aim is to understand the function of the *Cdx* factors in the regulation of other genes involved in the crypt villus axis differentiation. Based on the expression of the *Cdx1* and *Cdx2* factors in the crypt villus axis of the intestinal epithelium, and the expression pattern showed by the *Apc* in the intestine, our aim was to study if the *Cdx* factors are involved in the regulation of *Apc*. The hypothesis is that *Cdx1*, a gene expressed in the crypt involved in proliferation, would regulate in a negative manner the expression of *Apc,* in order to allow β-catenin to enter into the nucleus and activate *Cdx1*. Second, *Cdx2*, which is expressed in the villus and is

involved in differentiation of cells, may regulate *Apc* in a positive way; by activating *Apc*, β-catenin would be degradated, and as a consequence the cells would undergo differentiation.

Finding the mechanisms by which these genes are regulated will increase our understanding of the regulation of genes that play a key role during development and adult stages. Although this work aims to characterize two specific genes, the results will also add new knowledge to the role and function of conserved non- coding regions in the regulation of genes.

Chapter Two

Materials and Methods

2.1 Materials

2.1.1 Chemicals

Standard chemicals were obtained from Sigma (Dorset, UK) or BDH (Oxford, UK). Analytical grade Bacto agar, tryptone and yeast extract were obtained from Difco (Oxford, UK).

2.1.2 Enzymes

The Klenow enzyme, restriction enzymes, T4 DNA ligase and calf intestinal phosphatase (CIP) came from Boehringer Mannheim (Ingelheim, Germany), Amersham (Buckinghamshire UK), Cambio (Cambs, UK) or Pharmacia (New York, USA). Restriction enzymes were used at a concentration of 1 unit per μg of DNA in the buffer as recommend by the manufacturer.

2.1.3 Agarose gels

Restriction digests of DNA were resolved on a 0.8-2.0% agarose gel in Tris borate EDTA (10X TBE- 108 g Tris base, 55g boric acid, 40ml 0.5M EDTA per litre) buffer. Fragments prepared for ligations in Tris–acetate EDTA (50X TAE-242 g Tris base, 57.1 ml glacial acetic acid, 100ml 0.5M EDTA per litre) buffer (Sambrook et al. 1989).

DNA markers used were *BstEII* digested lambda DNA (700bp-8Kb), or *HpaII* digested Blue Script DNA (100bp- 700bp), 10X Ficoll loading buffer (25% Ficoll 400, 0.25% bromophenol blue, 0.25% xylene cyanol FF).

2.1.4 Bacterial strains

The following strain of *E coli* was used for cloning and expression studies. Strain: DH5α Genotype: *sup*E44Δ*lac*U169 (ϕ80 *lac*ZΔM15) *hsd*R17 *rec*A1 *end*A1 *gyr*A96 *thi*-1 *rel*A1.

2.2 Methods

2.2.1 Preparation of DNA fragments for ligation

DNA fragments for cloning were prepared from plasmids digested with excess of enzyme. The digests were electrophoresed on 0.8% agarose gels in 1X Tris-Acetate EDTA buffer (TAE). The fragment of interest was isolated by excising the band from the gel. The fragment of gel was placed in 300 µl of TAE in a dialysis membrane. The DNA was electroeluted into the buffer; the DNA was collected and extracted with phenol chloroform, and precipitated with 1/10 volume of 4M LiCl and 2.5 volumes of ethanol. The DNA pellet was dissolved in Milli Q water.

Fragments with 5' overhangs were filled in by Klenow in a volume of 10µl containing DNA, 1µl Klenow, 1µl 10X Klenow buffer and 0.2mM dNTPs and were incubated at 37 ^{0}C for 30min.

2.2.2 Dephosphorylation of vectors

The vector ends were dephosphorylated with Calf Intestinal Phosphatase (CIP) in a volume of 15 µl, containing 1µl CIP, 1.5µl 10X CIP buffer. For dephosphorylation of blunt ended fragments, a second incubation of 30min. was performed after the addition of 1µl CIP. The CIP was heat inactivated at 75^{0}C for 5min.

2.2.3 Ligations

Both blunt and sticky ended ligations were carried out for 16 hrs. at 16^{0}C. Ligations were performed in a volume of 10µl containing vector and insert, 1µl T4 DNA ligase, 1µl 10X ligation buffer.

2.2.4 Competent cells

Competent *E. coli* cells (DH5α) for transformations were prepared using TSS (10% Polyethylene glycol (PEG) 6,000, 5% Dimethyl sulphoxide (DMSO), 25mM $MgCl_2$). 2ml of an overnight culture was inoculated into 100ml LB media (Luria Bertani medium: 1% tryptone, 1% NaCl, 0.5% yeast extract) and incubated at 37^{0}C with shaking until OD_{600} reached 0.3-0.4. The cells were centrifuged in pre-cooled 50ml polypropylene tubes at 3, 500 rpm for 10min. at 4^{0}C. The pellet was

resuspended at one-tenth the original volume in ice cold TSS, and stored frozen at -70^{0}C. (Chung *et al.* 1989).

2.2.5 Transformations

Frozen competent cells were thawed, and then left on ice for ten minutes. Aliquots (100ul for a 10ul ligation) were dispensed into pre-cooled Eppendorf tubes, DNA was added, and incubated on ice for 30min. The cells were heated and shocked at 42^{0}C for 90s, returned to ice for one minute after the addition of 2-3 volumes of LB media. The transformed cells were incubated at 37^{0}C with shaking for 45 min., and then pelleted at 13,000 rpm for 5 min. (microfuge) and re-suspended in 50μl of LB. The suspension was plated on agar plates (60μg/ml ampicillin) and incubated overnight at 37^{0}C.

2.2.6 Plasmid mini- preps

Single colonies of transformants were inoculated into 3ml of LB media (70μg/ml ampicillin), and grown overnight at 37 ^{0}C with shaking. Cells (1.5ml) were pelleted at 13,000 rpm for one minute in Eppendorf tubes, and resuspended in 50μl Tris and Sodium buffer EDTA (TNE) (10mM Tris HCl pH8, 100mM NaCl, 1mM EDTA). 50μl phenol/chloroform isoamyl alcohol (25:24:1) was added, and mixed thoroughly. This was centrifuged at 13,000 rpm for 5 min. 50μl of the aqueous phase was transferred to a new tube and the DNA precipitated by adding 25μl 5M ammonium acetate and 170μl ethanol and incubated at 20^{0}C for 15min. The DNA was collected at 13,000-rpm for 10min. This was washed in 70% ethanol, briefly dried and dissolved in 20μl Tris- EDTA (TE) (10mM Tris HCl, pH 7.5 and 1mM EDTA, pH 8.0)

2.2.6 Large scale plasmid prep

Single colonies were inoculated into 2ml of LB media (70μl/ml ampicillin) and grown overnight at 37^{0}C. This was used to inoculate 500ml of LB media (70μl/ml ampicillin) and grown overnight at 37^{0}C with shaking. Cells were pelleted (GSA rotor) at 5,000 rpm for 15min. at 4^{0}C. The supernatant was drained and the pellet was re-suspended in 9ml solution 1 (50mM glucose, 25mM Tris, 10mM EDTA pH 7.), then incubated on ice for 5min., followed by the addition of 18ml of solution 2 (0.2 N NaOH, 1% SDS) and incubation for 10min. on ice. To the cell lysate 9ml of solution

3 (3M potassium and 5M acetate) was added and mixed by gentle swirling and a head over heels motion. Cell debris was pelleted at 8,000 rpm for 20min. at 4^0C. To the supernatant containing the DNA, 0.6 volumes of isopropyl alcohol was added, and the DNA pelleted at 10,000 rpm for 20min. at 4^0C. This was washed with 70% ethanol and dried briefly.

The DNA was either a) dissolved in 2ml of TE + 100μg/ml RNase and left for ½ hr at 37^0C, extracted with phenol/chloroform and precipitated by the addition of 2.5 volumes of ethanol, or b) dissolved in 5ml of TE and 0.5ml of ethidium bromide (10mg/ml) with 5.5g CsCl and centrifuged at 60,000 rpm for 20hr in a heat sealed polymer tube (Beckman L5-65 ultracentrifuge and Ti70.1 rotor). The supercoiled DNA was collected, and the ethidium bromide removed by repeated extraction with water plus salt-saturated n-butanol. The DNA was diluted 1:1 with TE and precipitated with 2.5 volumes of ethanol. The latter method was used when pure plasmid DNA was required for transfections of cells.

2.2.8 Restriction digestion

DNA from plasmid mini preps (0.5-1μg) was digested with the appropriate restriction enzyme in a total volume of 20μl (2μl 10X buffer, 1μl enzyme- 8 to 10u and 16μl water) with the first enzyme. After 3 hours the second enzyme was added if required and volume was made up to 40μl (4μl 10X buffer, 2μl enzyme- 8 to 10u and 24μl water). After 3 hours of digestion with the second enzyme, 1μl (20μg) of RNAse was added and samples were checked on agarose gels. Digestions were incubated at 37°C.

2.2.9 Preparation of DNA for sequencing

Double strand DNA for sequencing was purified on a caesium chloride gradient as described. Further purification was performed using a Micro Spin S-300 column (Amersham). The column was vortexed to re-suspend the resin and centrifuged at 3rpm for 2min. 50μl of the sample was loaded into the column and centrifuged at 3rpm for 2min. An aliquot of the eluent containing the pure DNA was checked on a 0.8% agarose gel and quantified. 300-500ng of template DNA with 5-10 pmol of primer in a total volume of 6μl were used for sequencing.

2.2.10 DNA sequencing

Sequencing was performed by automated cycle sequencing ABI sequencer. Nucleic acid sequence homology searches were performed using the FASTA programs of the GCG sequence analysis package and the GeneBank™/EMBL database (www.ebi.ac.uk/embl/index.html).

2.2.11 Polymerase Chain Reaction

PCR mixtures contained 200μM of each dNTP, 50pmol of each primer, 50ng of template DNA, 5 units of *Taq* polymerase (Roche), 10mM Tris-HC, pH 8.3, 50mM KCl, 1.5mM $MgCl_2$, in a total volume of 50μl overlaid with mineral oil. The PCRs were performed in a Thermal Cycler from MJ Research Inc (PTC-100).

2.3 Cell culture methods

2.3.1 Cell lines

The cell lines used on transfection and transactivation assays were CaCo2 and IEC-6 cells. CaCo2 cells or Caucasian colon carcinoma cells are colon cancer cells from human origin that display very similar characteristics to intestinal enterocytes. These cells can be grown to differentiation very quickly; so comparative studies on gene expression can be performed. IEC-6 cells are normal intestine cells from rat origin, which display all the characteristics of intestinal epithelial cells and do not express *Cdx1* and *Cdx2* genes.

2.3.2 Cell cultures

All cells were cultured at 37^0C and 5% CO_2. All media components were purchased from Gibco Invitrogen: Dulbeccos Modified Eagle Media (DMEM), non-essential amino acids and the foetal calf serum provided by Labtech Int. were purchased sterile or filter sterilised through a 0.2μm filter. Trypsination of cells for expansion or freezing was performed by aspirating the culture medium, washing the monolayer with PBS, adding 3-5ml of trypsin (Gibco) and immediately aspirating to leave a thin layer of trypsin covering the cells. The cells were incubated at 37^0C until the cells detached from the surface, fresh medium was added and the cells re-suspended with a pipette.

2.3.3 Freezing cells

All cell lines were stored in liquid nitrogen as single cell suspensions in 40% DMEM, 50% FCS, 10% DMSO (freezing medium) in cryovials. The cells were trypsinised, and then pelleted by centrifugation at 1,500 rpm for 5min. The cell pellet was re-suspended in freezing medium and the cryovials were frozen overnight in dry ice and then stored under liquid nitrogen until use.

2.3.4 Transfection by Fugene™ reagent

The day before transfection, cells were plated at 1-3X10^5 cells into 30mm dishes at low density (10-15% confluency). The cells were at 50-80% confluency on the day of the transfection. One hour prior to transfection the medium was changed. The Fugene™ procedure was used to transfect cells. In a labelled tube, 97µl of serum free media was added and incubated for 5 min. at room temperature. To a second set of labelled tubes the required amount of DNA was added. The mix of Fugene™ and media was added to the DNA and incubated at room temperature for 15 minutes. The resulting mixture was added to the plates.

Chloramphenicol acetyl transferase (CAT) reporter plasmids (equimolar amount of reporter constructs and empty vector) were co-transfected with 0.5µg of CMV Cdx1 or CMV Cdx2. 0.5µg of p27 LacZ or pNLS LacZ was used as internal control to normalise for variations in transfection efficiency. Bluescript II SK+ plasmid (Stratagene) was used to adjust the total amount of transfected DNA to 1-1.5µg. The cells were incubated for 48 hrs at 37^{0}C and 5% CO_2.

2.3.5 Staining for β-galactosidase

Cells were washed twice with ice-cold PBS pH7.4 to remove dead cells, and then fixed on ice for 3.5 min. in fixative containing 2% formaldehyde/0.2% gluteraldehyde/PBS. The cells were then washed with PBS pH7.4 and incubated in substrate solution (1mg/ml X-gal- Boehringer – 5mM potassium hexacyanoferrite, 5mM potassium hexacyanoferrate, 2mM $MgCl_2$), at 37^{0}C. Staining was carried out from 1-24 hours. The stained cells were washed with PBS and stored at 37^{0}C in 10% glycerol/PBS. The number of blue cells was counted to check for transfection efficiency.

2.3.6 Protein extraction

For β-gal and CAT assays, protein extracts were prepared from transfected cells. The medium was removed from the samples, which were then washed three times with ice-cold PBS pH7.4 to remove dead cells. 0.5ml of 1X lysis buffer was added to the samples and incubated with stirring for 30min. The samples were collected in pre-cooled eppendorf tubes and centrifuged at 4^0C during 10min. The aqueous phase was transferred in 250μl aliquots into pre- cooled eppendorf tubes and stored at -70^0C.

2.3.7 Protein estimation

Protein estimation was performed using the BCA Protein Assay Reagent Kit from Pierce; protein was estimated either in test tubes or in micro- plates. For the first method 0.1ml of each replicate standard curve or unknown sample was loaded into its own-labelled test tube and 2.0ml of Working Reagent (WR) was added to each tube and mixed. Samples were incubated at 73^0C for 30min. Absorbance was measured in a spectrophotometer at 562nm. Standards were prepared using bovine serum albumin (BSA) 2.0mg/ml. For the micro-plate method, 25 μl of each standard or unknown solution was loaded into the microplate well and 200μl of WR was added to each well. The plate was covered and incubated for 30minutes. Absorbance was measured in a spectrophotometer at 562nm.

2.3.8 β-gal and CAT ELISA assays

Promoter activity in transfected mammalian cells is generally studied by linking the promoter sequence to a bacterial gene encoding a detectable reporter protein such as CAT or β-galactosidase (β-gal). These two reporter genes have become standard markers used to measure transfection in most cell lines. The bacterial CAT, having no eukaryotic equivalent, has become a useful marker used in transfection experiments with eukaryotic cells. In the case of β-gal the system used can differentiate between endogenous lysosomal β-gal activity and the transfected bacterial enzyme.

2.3.9 β-galactosidase ELISA assays

β-gal enzyme-linked immunosorbent assay (ELISA) assays were performed using the β- galactosidase ELISA kit (Roche) for quantitative determination of β-gal from *E coli* in transfected eukaryotic cells. Each transfection experiment was carried out in duplicate or triplicate.

2.3.10 CAT ELISA assays

CAT ELISA assays were performed using the CAT ELISA colorimetric enzyme kit (Roche) for quantitative determination of CAT from *E coli* in transfected eukaryotic cells. All CAT ELISA values were normalised for transfection efficiency by calculating the ratio of CAT activity to β-gal in each transfected plate. Each transfection experiment was carried out in duplicate or triplicate.

2.3.11 Characterisation of the APC regulatory region

CAT activity of the promoter constructs was measured relative to the promoterless pCAT BASIC. pCAT CONTROL (Promega), containing SV40 promoter and enhancer sequences, was used as a positive control. A 607bp fragment spanning positions –290 to +317 of *APC* was subcloned, in the correct orientation, upstream of the CAT gene to generate pApcSP (Wedgwood *et al.* 2000).

2.4 Resources available at the MRC UK HGMP Resource Centre

2.4.1 The Fugu genomic clone libraries

The *Fugu* cosmid and BAC libraries have inserts that range from approximately 35 to 45Kb (average 40Kb) and 60 to 120Kb (average 80Kb), respectively. The vectors used are Lawrist 4 for cosmids and pBeloBACII for BACs. All *Fugu* clones are available from MRC geneservice (http://www.hgmp.mrc.ac.uk/ gene service/index.shtml). The initial clones sequenced during this project were cosmids, as these were identified in the *Fugu* Landmark Project (http://fugu.hgmp.mrc.ac.uk/) as containing the genes of interest. In July 1999, a

Fugu BAC library was produced by Incyte Genomics. These genomic clones replaced the cosmids in the analysis due to their larger insert size.

2.4.2 The Fugu BAC library

The *Fugu* BAC library was prepared at Incyte using *Fugu* genomic DNA supplied by Dr. Greg Elgar (MRC UK HGMP Resource Centre). The library contains 42,624 clones arrayed in 111 (384-well) plates. The library plates are numbered from 176-286. The *Fugu* group at the MRC UK HGMP Resource Centre prepared high-density filters of the library. The library is equivalent to approximately 10X coverage of the entire *Fugu* genome.

2.4.3 Primer synthesis and usage

Primers were synthesised by the Sanger Centre Synthesis Group (Wellcome Trust Genome Campus, Hinxton Hall, Cambridge). These were supplied in ammonia that was removed by centrifugal vacuum evaporation (Jouan centrifugal evaporator RCT90-RC1022) for 45 minutes. Primers were quantified by measuring their optical density (OD) at 260 nm. Their concentration was calculated using the equation: Concentration in mM = [(O.D. at 260nm) (20) (dilution factor)] / [(325) (oligo length in bp)]. The value of twenty was used as the single stranded DNA constant (Sambrook *et al.*, 1989). A working stock of the primer was made at a concentration of 10 μM.

2.4.4 General PCR conditions

General polimerase chain reaction (PCR) reactions were usually carried out in a 25 μl reaction volume using: 5 μl 10X buffer, 5 μl 2mM dNTPs, 1.5 μl 50mM $MgCl_2$, 26.5 μl double distilled water, 1 μl Taq polymerase (Bioline), 1μl DNA (20μl/ml) and 5 μl per primer pair. The reactions were denatured at 95 °C for 30 seconds, annealed at 55°C to 65°C for 30 seconds (according to the Tm of the primers being used), and extended at 72°C for 60 seconds with a total of 30 to 35 cycles. A standard initial 95 °C for 2 minutes denaturation step and a final 72°C for 7 minutes elongation step were also used for the PCR reactions.

The PCR reactions were also conducted with limited dilutions of dNTPs and vector primers (an eighth of the normal concentration of dNTPs and a fourth of the

normal concentration of primers). These conditions allowed the direct sequencing of PCR products without the requirement of product purification, and also reduced the number of non-specific bands obtained during the amplification of DNA.

2.4.5 Reverse transcription polymerase chain reaction (RT-PCR)

RT-PCR was used to confirm the exon/intron boundaries of the identified genes in *Fugu*, as well as to examine the expression patterns. 1 μg of total RNA was used for cDNA synthesis (using the RNA stocks available from the *Fugu* Genomics group). The genomic contamination in the total RNA stocks was removed using DNAse treatment. The cDNA synthesis was carried out using the reverse transcriptase (RT) kit (Promega) following the manufacturer's instructions. Using actin primers after cDNA strand synthesis assessed the quantity and quality of the cDNA obtained. The actin primers used included actin forward (5'acagactacctcatgaagatcct3') and actin reverse (5'gaggccaggatggagcctcc3'). The primers were specifically designed to amplify *Fugu* actin sequences that span an intron, and therefore were used to detect genomic contamination of the cDNA. The design of the primers was based on published *Fugu* actin sequences (Venkatesh *et al.* 1996).

The PCR was carried out in a 25 μl reaction, consisting of 1 μl of cDNA with limited dilutions of dNTPs and primers, repeated 25 times with the following PCR cycles: initial denaturation at 96°C for 2 minutes only, 95°C for 30 seconds, 58°C for 30 seconds, 72°C for 1 minute, and a final extension of 72°C for 5 minutes. When PCR products were difficult to obtain, PCR enhancing agents were used, namely 2% DMSO, which was added during the first and second round of PCR. DMSO optimises the PCR amplification by facilitating strand separation by disrupting base pairing.

2.4.6 Identification of the 5' end of genes

5' Rapid Amplification of cDNA Ends (RACE) allowed the amplification of unknown sequences at the 5' end of the mRNA. The BD SMART^{TM} RACE cDNA amplification kit (BD Biosciences Clontech) was used in an attempt to obtain the 5' end of genes. The principle of the 5'RACE consists first, in the synthesis of cDNA from total or polyA+ RNA using a modified oligo (dT) primer, termed the 5'-CDS primer: 5'- agg cag tgg tat caa cgc aga gta $c(t)_{25}$ V N-3' ,the BD SMART II A

oligonucleotide 5'- aag cag tgg tat caa cgc aga gta cgc ggg -3' and the BD RT. Then, a 5'RACE PCR is carried out using gene-specific primer and the long universal primer: 5'- cta ata cga ctc act ata ggg caa gca gtg gta tca acg cag agt -3'. A second 5'RACE PCR is performed using the short universal primer 5'- cta ata cga ctc act ata ggg c -3'and a nested specific primer. The final 5'RACE PCR product is then purified, cloned and sequenced.

The first strand cDNA synthesis was done using 1μg of total RNA extracted from *Fugu* gut tissue, 1μl 5'-CDS primer and 1μl BD Smart II A oligo, incubated 70^{0}C for 2min. followed by 2min. at 4^{0}C; then 2μl 5X buffer, 1μl DTT (20mM), 1μl dNTPs (10mM) and 1μg BD reverse transcriptase were added to a final vol of 10μl; samples were incubated at 42^{0}C for 1.5hrs. 100μl of Tricine-EDTA buffer were added to the reaction.

The first 5'RACE PCR reaction was prepared adding 2.5μl of the cDNA reaction, 5μl universal primer mix (UPM) (10X), 1μl gene specific primer (GSP1) (10mM)[5' ggcc cgg cga ttc tgg aac cag atc 3'], 5μl 10X 2 PCR buffer, 1μl dNTPs (10mM), 1μl BD polymerase mix and 34.5μl water. The PCR program was 94^{0}C for 1min.; 94^{0}C 30sec, 72^{0}C for 3min. during 5 cycles; 94^{0}C for 30sec, 70^{0}C for 30sec, 720C for 3min. during 5 cycles the final cycling was at 940C for 30sec, 68^{0}C for 30sec, 72^{0}C for 3min. during 27 cycles followed by 72^{0}C for 5min. For the second 5' PCR reaction, 5μl of the first PCR product were diluted into 245μl of Tricine EDTA buffer; 5μl of the diluted primary PCR were used as a template, using 1μl NUP (10X) and 1μl of the NGSP (10mM)[5' aggc tga ggg ctg tgg cca gct ccg 3'], the remaining components were kept as the first reaction. A positive control was prepared using the human placental total RNA provided for the kit, which was prepared exactly in the same conditions as the sample.

2.4.7 DNAse treatment of RNA

Total RNA from different tissues was treated with DNAse to remove genomic contamination. A mixture of 120 μl 100mM $MgCl_2$/10mM DTT, 2.4 μl DNAse I, 1.2 μl RNAse inhibitor and TE buffer was added to a 1 μg RNA sample, to make a final volume of 700 μl. The samples were incubated at 37 °C for 3 hours. RNA was then extracted by adding 50 μl of phenol:chloroform:isoamyl alcohol (25:24:1), vortexed and centrifuged for 1 minute. The top aqueous layer was removed to a fresh

eppendorf. 75 µl of chloroform was added, vortexed and centrifuged for 1 minute. The top layer containing the RNA was pipetted into a fresh eppendorf, precipitated with 300 µl 25:1 K Acetate/ethanol and then washed twice with 70% (v/v) ethanol. The resulting pellet was air-dried, resuspended in 100 µl of double distilled water, and used directly during RT-PCR.

2.4.8 Cloning PCR products into vectors

PCR products were cloned into the pGEM ^{R} -T Easy Vector System (Promega) or pBluescript vector prior to sequencing. The pBluescript vector (Statagene) used for cloning PCR products did not have the 5'-phosphate group removed by dephosphorylation from the vector after being linearised. This is because PCR products do not have a 5'-phosphate group, which when provided by the vector, is needed by the enzyme T4 DNA ligase 400 (U/µl) during the ligation reaction.

2.4.9 Cloning into the pGEM Easy Vector

The pGEM ^{R} -T Easy Vector System (Promega) allowed the cloning of PCR products prior to sequencing. The vectors contain T7 and SP6 RNA polymerase promoters flanking a multiple cloning region within the α-peptide coding region of the enzyme β-galactosidase (used for colour screening of recombinant clones). The PCR products were ligated into the pGEM vector, following the manufacturer's instructions, and PCR amplified with T7 short (5' taatacgactcactatagg 3') and SP6 short (5' gatttaggtgacactatag 3') primers. PCR products were then sequenced using T7 long (5' taatacgactcactatagggcga 3') or SP6 long (5' gatttaggtgacactatagaatac 3') primers.

2.5 Bioinformatics tools and programs

2.5.1 BLAST

The Basic Local Alignment Search Tool, or BLAST program (Altschul *et al.* 1997) enables the rapid identification in a large database of regions with similarity to a given query sequence. The process works by using local alignment algorithms that attempt to isolate regions in sequence pairs that have high levels of similarity. Although BLAST does not guarantee to find the best local alignment, it is the most

effective heuristic search method available, in particular when used against large data sets.

2.5.2 NIX- a nucleotide identification program

Once the genomic sequence around the region of interest in *Fugu* was produced, the next step involved the process of finding and identifying genes. A nucleotide identification program (NIX) (http://www.hgmp.mrc.ac.uk/Registered/Webapp/NIX/) is an integrated automated DNA sequence analysis tool. It masks the sequence for repeat regions, carries out BLAST searches against databases, and also runs *ab initio* exon-finding programs, displaying all the results graphically. Viewing all the results in conjunction offers a collective way of interpreting the predictions and BLAST matches of the genomic sequence being analysed, facilitating the identification of exons and prediction of proteins. The range of programs displayed in NIX include: Eponine, GRAIL, TSSW, GENSCAN and Fgenes for promoter predictions; FEX, Hexon, MZEF and Genemark for exon predictions; GRAIL, Genefinder, Fgene, GENSCAN, Fgenes and HMMGene for gene predicitions; and finally GENSCAN, Fgenes and GRAIL for polyA predictions. BLAST sequence similarity searches against unigene, mRNA, EST, EMBL, *E. coli* and vectors, as well as RepeatMasker, are also included.

2.5.3 Internet resources

A range of websites was used regularly during this project. The European Molecular Biology Laboratory (EMBL) represents Europe's primary nucleotide sequence database. The database contains a collection of DNA and RNA sequences, with new entries being updated daily. EMBL was used to retrieve orthologous or paralogous sequences to be used in the comparative analyses. The Ensembl website generates and maintains annotation for eukaryotic genomes, including human, mouse, zebrafish, mosquito and *Fugu*, and has been a major source in the annotation of the Human Genome Project draft sequence. The National Centre for Biotechnology Information (NCBI) is a central internet resource that offers a general repository for molecular biology. Links within the main website include: online mendelian inheritance in man (OMIM), map viewer, BLAST, links to literature (Pubmed), Entrez data-mining tools, the cancer genome anatomy project, gene annotations, and

the LocusLink gene viewer. SRS6, produced by Lyon Bioscience AG, allowed the retrieval of sequences deposited in the public databases.

The website addresses are shown below:

NCBI: http://www.ncbi.nlm.nih.gov/

Ensembl: http://www.ensembl.org/

SRS: http://srs.hgmp.mrc.ac.uk/srs6/

EMBL: http://www.ebi.ac.uk/embl/

2.5.4 Emboss

DNA sequences were manipulated using EMBOSS, a sequence analysis tool (Rice *et al.* 2000). EMBOSS is a free open source software analysis package encompassing over 100 programs. The programs used during this project included sequence alignments, rapid database searching and nucleotide sequence pattern analysis. EMBOSS was used to translate *Fugu* genes and deduce their putative protein sequences.

2.5.5 ClustalX

ClustalX (Thompson *et al.* 1994) generates multiple sequence alignments and displays the results in a graphical interface. Although the default parameters were used for all the alignments during this project, these can be changed accordingly. The multiple sequence alignment parameters used included: Gap opening penalty 10.00, Gap extension penalty 0.20 and DNA transition weight of 0.5, using a Gonnet 250 protein matrix.

2.5.6 Mlagan

Mlagan is a global aligment tool kit (Brudno *et al.* 2003) used to do pairwise and multiple alignment of large scale sequences. Mlagan works by first creating a local alignment of sequences, it is that genes or sequences with very high homology present in the complete sequence are aligned first, then a rough global map is created in base to the local alignments and finally, a computation of a final global alignment is done using the best alignment from the rough global map. The final Mlagan

alignment is visualized as Vista plot, which allows and easy identification of the conserved sequence in the complete alignment.

Mlagan was used to look for the conserved non-coding sequences surrounding the genes of study; sequence conservation was measure using a 40 bp window and a 60% identity. The *Fugu* sequence was in most of the cases used as base line sequence.

2.5.7 Theatre

Theatre is a software tool designed for the comparative analysis of genomic sequences. Theatre predicts positions of coding regions, repetitive sequences and transcription factors binding sites present in the studied sequence. Theatre incorporates a repertoire of programs required for the complete analysis of the sequence. It uses Clustal W V.1.74 to perform an alignment of sequences, Blast V. 2.2.3 and Gene Mark V. 2.3 to identify protein coding regions and open reading frames, MatInspector V.2.2 and Tfscan (EMBOSS) V.2.5.1 to predict transcription factors binding sites and CpG plot (EMBOSS) V.2.5.1 and RepeatMasker to identify repeats in the sequence. Theatre provides a pairwise or multiple alignment of sequences with the transcription factor binding sites present and conserved in the aligment (Edwards *et al.* 2003).

2.6 Fish maintenance and embryo injection

Zebrafish were raised and bred at 28.5^{0}C. Embryos were staged accordingly with the number of hours after fertilization (hpf). Embryos were injected using 100ng/μl of coiled DNA and 0.1% of phenol red. Injections were done into zebrafish embryos produced by natural mating between the one to eight cells stage using an Eppendorf (Hamburg, Germany) FemtoJet pressure injection system. Injected embryos expressing GFP were anaesthetized in Tricaine and analyzed using an Olympus (Tokyo, Japan) IX81 motorized inverted microscope. Images were captured using an FVII CCD monochrome digital camera and analySIS image-processing software. For embryos injected with *LacZ*, embryos were fixed at the desired stage and stained for *LacZ* activity.

2.7 X-gal staining and histological analysis of embryos

β-gal staining allows identification of embryonic tissues/cells expressing *lacZ* marker protein by development of pigmented (blue) product in the presence of *lacZ* enzymatic activity.

Transgenic embryos were fixed in 0.1% glutaraldehyde–PBS for 5 min., followed by three washes with PBS for 5 min. at room temperature. Staining was carried out for 5–12 h at 37°C in PBS containing 1 mg/ml X-gal, 5 mM $K_3Fe(CN)_6$, 5 mM$K_4Fe(CN)_6$, and 2 mM $MgCl_2$. Stained embryos were washed twice with PBS and immediately stored in 0.1% glutaraldehyde in PBS. Embryos were washed with PBS (3 changes of 5 min. each, followed by incubation of 30 min. each in 30%, 50% and 70% ethanol, and stored in ethanol at -20^0C.

Chapter Three

Identification of the *Cdx* genes in *Fugu rubripes*

3.1 Introduction

The homeobox genes are transcription factors involved in the determination and sculpture of the body plan in early development. One of the main characteristics of these genes is a 180bp region named the homeobox, which codes for the homeodomain in the protein. This homeodomain is highly conserved across species, and due to this level of conservation, several homeobox genes from different organisms have been identified to date. The *Cad* genes are homeobox genes. The *Drosophila Caudal* gene was the first identified by its homology with the *Ant* homeobox transciption factors (Mlodzik *et al.* 1985).

The murine *Cdx1* was the first vertebrate *Caudal* homeobox gene to be isolated using a *Drosophila Cad* probe from a 8.5 dpc mouse embryonic cDNA library. It is a 1.8Kb transcript that encodes a 268aa sequence (Hu *et al.* 1993). The murine *Cdx2* was isolated using a cDNA library made from mRNA of colonic crypts from adult mouse and a fragment with the *Cdx2* homeodomain as a probe based in the *Drosophila* homeodomain sequence. The murine *Cdx2* is a 1.8Kb transcript, which translates into a 311aa. protein (James *et al.* 1994). The murine *Cdx4* was isolated using an 8.5 dpc mouse cDNA library and two thirds of the homeobox as a probe. The *Cdx4* transcript is the smallest of the murine *Cdx* genes with 856bp length and encoding 282aa. protein (Gamer and Wright 1993). There is a high degree of conservation of the gene sequence between various species. The *Cdx* genes contain three coding exons and two introns. The first intron is normally very long compared with the second intron. The homeobox is usually contained between the end of the second exon and the first part of the third exon. The homeobox codes for 60aa. peptide, the homeodomain; upstream of the homeodomain there is a hexapetide sequence (HEX) which is used to classify the *Caudal* -like genes (Burglin *et al.* 1989).

The human *Cdx* genes display a high level of sequence identity with the mouse *Cdx* genes. The nucleotide and amino acid sequence, the genomic structure and the map position of the locus are also conserved. The three human *Cdx* genes also contain three coding exons and two introns; the homeobox is also located between the

second and third exon and the hexapeptide domain which is separated from the homeobox by the first intron is also maintained in mouse and human *Cdx* genes (Bonner *et al.* 1995).

Not much is known about the biochemistry of the Cdx proteins. There are well conserved motifs in the proteins across the family and across species; the proline rich regions present in the sequence have been suggested to act in the transactivation and oligomerisation of the protein. The presence of phosphorylation and acetylation sites are also characteristics of the *Caudal* proteins; the phosphorylation site is thought to be involved in DNA binding along with the homedomain of the protein.

Given the high conservation of these genes across species, it is likely that the elements involved in the regulation of Cdx1 are also conserved across species. To identify the conserved regulatory elements involved in the regulation of the *Cdx1* gene, I reverted to the use of comparative genomics and the *Fugu* genome. The primary idea was to identify the *Fugu Cdx1* gene in the pufferfish genome, and based on gene conservation, I aimed to look for the conserved non- coding sequences present in the upstream, downstream and intronic regions that might be candidates for the regulatory elements of the *Cdx1* across species.

3.2 Materials and methods

3.2.1 Expression of *Fugu Cdx* by RT-PCR

Total RNA from wild type adult pufferfish was obtained from Dr. Greg Elgar (MRC UK HGMP Resource Centre). Primers for *frCdx1* expression, forward: 5'-ATCCCAGGCCCCTATGA -3' and reverse: 5'-caggaaagcatcagacc -3'. PrimerS for *frCdx2* expression, forward: 5'-ggactaaagacaagtaccgg -3', reverse: 5'-gaagatctggttccagaatc -3' and primers for the *frCdx4* expression, forward: 5'-ggaccagtaagaagatcaag -3' reverse: 5'-gcagagctaaagagaggaag -3' Internal primers for the β-actin gene were designed to confirm that there was no contamination of genomic DNA. The sequence of the *β-actin* primers were: forward primer 5'-tacagactacctcatgaagatcc-3' and the reverse primer 5'-gaggccag gatggagcctcc-3'. The PCR product size for the *β-actin* was 500pb.

The PCR was performed in 20µl containing 1µl cDNA product from the reverse transcription and 25µM of each primer, with the following program: annealing temp. 65^{0}C for 30s; extension temp. 72^{0}C for 40s; 35 cycles. The PCR product size for the *FrCdx1* was 910bp and the PCR product size.

3.3 Results

3.3.1 Identification of Fugu Cdx genes

This work was done subsequent to the completion of the *Fugu* genome sequencing. The *Fugu* genomics project database was used to search for the *Fugu Cdx* genes; all *Fugu* DNA sequences were extracted from this database. The genomic sequence for the *Cdx* genes for human and mouse were extracted from the Ensembl database. This section will describe how we identified the *Fugu Cdx* genes by homology using the mouse and human *Cdx* sequences from the Ensembl database. The mouse and human *Cdx* genes (including aa sequence, transcriptional organisation, coding sequences and intergenic distances) have been described and analysed previously, and we make use of this available information to characterize the *Fugu Cdx* genes in an improved manner.

To identify the *Cdx1* gene in *Fugu*, we extracted the aa sequence of the mouse *Cdx1* (NM_009880), and blasted it against the *Fugu* genome database. The blasted sequence hit three mayffolds, M001324, M000598 and M000580. Each mayffold was displayed using the NIX program, which is one of the features contained in the *Fugu* database used for the analysis of a single clone. The NIX program contains a package of exon-finding programs that predict the gene structure, gene sequence, aa sequence and gene location in a specific DNA sequence.

The NIX program showed that the three mayffolds contain a predicted *Caudal* gene. This suggested that *Fugu* contains the three *Cdx* genes present in mammals. However, assignation of the distinctive name and a complete characterisation of the *Fugu* Cdx genes were not available yet. The first attempt to assign a name to each *Caudal Fugu* gene was made using the aa sequence predicted by NIX. However, the prediction only showed the homeodomain, which is highly similar between the three

genes; because the gene structure is exactly the same for the three *Cdx* genes, we opted to use the linkage of the genes to distinguish among them.

Using the transcriptional organisation from mouse and human *Cdx1*, *Cdx2* and *Cdx4*, we assigned the *Fugu Cdx1* (*frCdx1*) to the mayffold 1324 (M001324); the *Fugu Cdx4* (*frCdx4*) was allocated to the mayffold 598 (M000598) and the *Fugu Cdx2* (*frCdx2*) was assigned to the mayffold 580 (M000580). The blast of the aa sequence of human CDX1 and mouse and human Cdx2 and Cdx4 hit the same mayfflods in the same order.

3.3.2 Transcriptional organisation of Fugu Cdx genes

3.3.2.1 Transcriptional organisation of *Fugu Cdx1*

Comparison of the genomic organisation of the *Fugu*, mouse and human *Cdx1* shows synteny conservation between the genes. A specific chromosome location for the *Fugu Cdx1* has not been assigned yet; the gene can be located in the mayffold 1324, which contains the complete genomic sequence of the gene plus the flanking regions of the gene.

At the 5' end of the *frCdx1*, the nearest gene is the beta platelet- derived growth factor receptor precursor gene (*Pdgfrβ*), located at 11.9Kb, followed by the macrophage colony stimulator factor I receptor precursor (*Mcsfr*), located 15Kb upstream of the *Fugu Cdx1*. At the 3' end, the nearest gene is the slow nerve growth factor receptor gene (*TrK-A*), located 953pb downstream of the *frCdx1* (Figure 3.1).

The human *CDX1* is located on chromosome 5. The beta platelet- derived growth factor receptor precursor gene (*PDGFRβ*) is the 5' proximal gene to the *CDX1* located at 11.13Kb; the second nearest gene is the macrophage colony stimulating factor I receptor precursor (*CSF1R*) situated 11.5Kb from the *CDX1* gene. In the 3' region, the nearest gene is the solute carrier family 6 (neurotransmitter- transporter, L-proline) gene (*SLC6A7*) placed at 6.2Kb, followed by the calcium/calmodulin-dependent protein kinase type I alpha chain (*CAMK2A*) located 8.4Kb from the *CDX1*.

The mouse *Cdx1* is positioned on chromosome 18; synteny is highly conserved between mouse and human in both the 5' and 3' regions. The nearest 5'

gene to the *Cdx1* is the *Pdgfrβ* located at 8.951Kb; the second proximal gene is the *Csf1r* located at 29.5Kb from the *Cdx1*. In the 3' end, the nearest gene to the *Cdx1* is the *mSlc6a7* placed at 4.66Kb; the second neighbouring gene at this end is the *Camk2a* situated 11.8Kb from the Cdx1 (Figure 3.1).

In contrast with the human and mouse *Cdx1* in which transcriptional orientation is well conserved, in the *frCdx1*, synteny conservation with mouse and human *Cdx1* is only present in the 5' end of the gene, where the *pdgfrβ* and the *Csf1r* (*Mcsfr* in the case of *Fugu*) genes are well preserved. However, conservation is broken at the 3' region of the gene.

The *mpdgfrβ* and *hPDGFRβ* contain a distinctive long intron between the 5'UTR and the 1st coding exon of the gene. The 5'UTR of the *Fugu pdgfrβ* has not been characterised yet. Comparison analysis of the upstream *Fugu pdgfrβ* upstream region did not show any sequence conservation with the human or mouse 5'UTRs. The distance between the *frCdx1* and the *frpdgfrβ* used here is from the 1st coding exon of both genes.

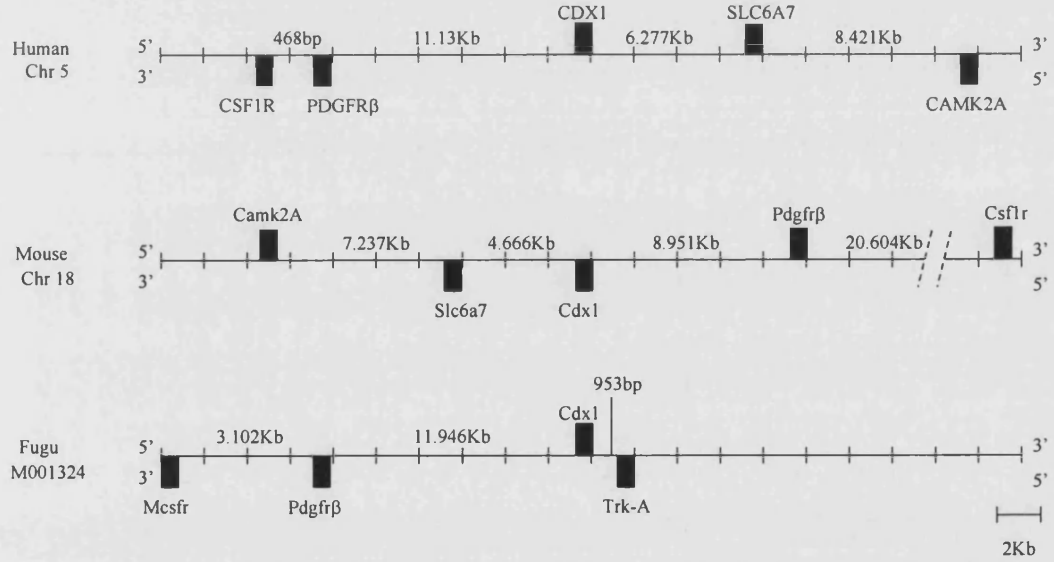


Figure. 3.1. Transcriptional orientation of the *Cdx1* gene in human mouse and *Fugu*. The *Cdx1* genomic regions from human and mouse were extracted from Ensembl. The genomic DNA extracted was up to the two nearest neighbouring genes. The *Fugu Cdx1* genomic region (M001324) was extracted from the *Fugu* database. Black boxes indicate the position of each gene through the chromosome; genes located above the line are transcribed 5'→3'; genes located

under the line are positioned in the reverse DNA strand and are transcribed in the opposite orientation. The scale bar for the human, mouse and *Fugu* is 2Kb.

3.3.2.2 Transcriptional organisation of *Fugu Cdx2*

The *Fugu Cdx2* was mapped in the mayffold 580 (M000580). This mayffold contains the complete *frCdx2* gene and the flanking regions to the nearest genes. In the 5' region, the two nearest genes to the *frcdx2* are the adapted related complex subunit gene (*ARCS*) and the arrestin β2 red cell isoform gene (*Carr*). The immediate 5' gene is *Carr*, located 4.5Kb away from the *frCdx2* followed by the *ARCS* placed 9.9Kb from the *frCdx2*. To the 3'end, the nearest gene is the ring finger protein 26 (*RFP*) situated at 10.7Kb; the second nearest gene is the insulin precursor protein (*IP*) located 13.2Kb from the *frCdx2* (Figure 3.2).

In the case of human and mouse *Cdx2*, the gene is located in chromosome 13 and 5 respectively; synteny is well conserved between these two species in the 5' end. The closest gene to the *Cdx2* is the cytokine receptor precursor (*Flt3*), which in the case of human is located at 34Kb and in the case of mouse is 23.9Kb from the *Cdx2*. The second nearest gene to the *Cdx2* is a predicted gene which in the case of the human as been assigned as *Q9H7FL*, and in the mouse as *Q9C2M6* located at 40.8Kb and 152.9Kb respectively from the *Cdx2*. Although no name or function to this gene has been assigned in mouse and human, the gene probably is not conserved between these two species, especially regarding the difference in the intergenic distance between the predicted gene and the *Cdx2* gene. To the 3' end, the gene immediatly next to *Cdx2* is the insulin promoter factor (*Ipf1*) situated 36.8Kb in human and 25Kb in mouse from the *Cdx2*. The second nearest gene, Gsh1encodes a homeobox protein, which locates 162.8Kb in the human and 105Kb in the mouse away from the *Cdx2* (Figure 3.2).

The transcriptional organisation of the *Fugu Cdx2* does not show any synteny conservation when compared with mouse and human. None of the neighbouring genes seems to be conserved, except for the predicted insulin precursor (*IP*) gene in *Fugu*, which is reminiscent of the *Ipf1* present in human and mouse. However, *Ipf1* is the closest gene in the 3' end in mouse and human, and *IP* is the second nearest gene located in the 3' end of the *frCdx2*.

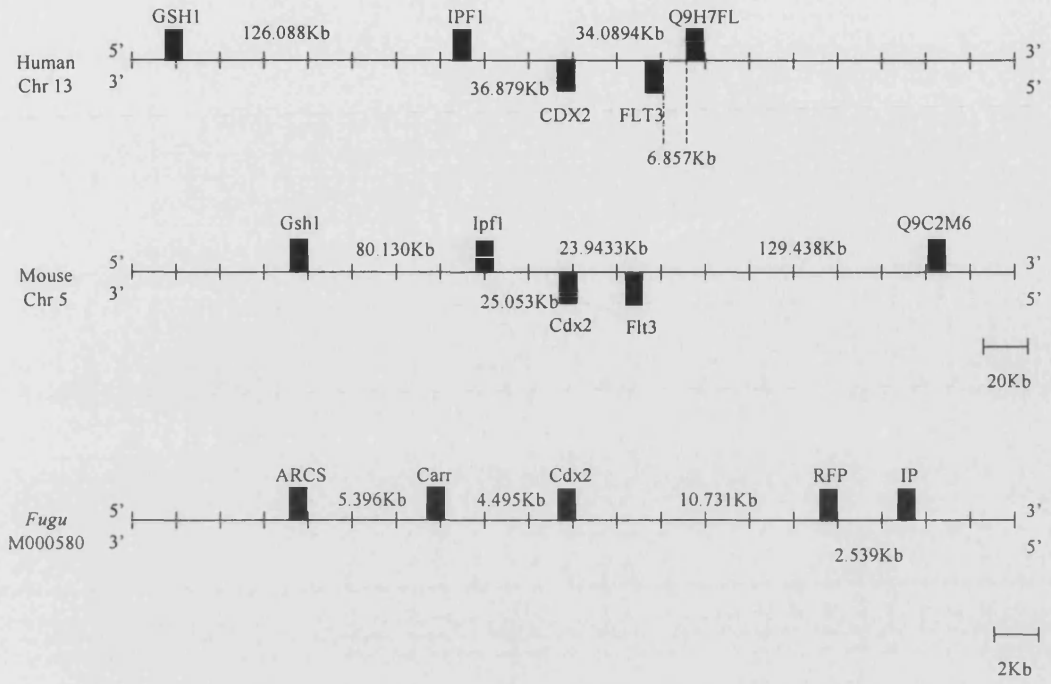


Figure. 3.2. Transcriptional orientation of the *Cdx2* gene in human mouse and *Fugu*. The *Cdx2* genomic regions from human and mouse were extracted from Ensembl. The genomic DNA extracted was up to the two nearest neighbouring genes. The *Fugu Cdx2* genomic region (M000580) was extracted from the *Fugu* database. Black boxes indicate the position of each gene through the chromosome, genes located above the line are transcribed 5'→3', genes located under the line are positioned in the reverse DNA strand and are transcribed in opposite orientation. The scale bar for the human and mouse is 20Kb.

3.3.2.3 Transcriptional organisation of *Fugu Cdx4*

A chromosomal location for *Fugu Cdx4* has not been assigned yet; the *frCdx4* was located in the Mayffold 598 (M000598), which contains the complete *frCdx4* and the neighbouring flanking genes. The 5' flanking gene is tyrosine kinase receptor (*FLK1*) at 7.3Kb; the second neighbouring gene in this end is the testis express gene (*TXS*) located 16.3Kb from the *frCdx4*. At the 3' end, *Chic1* is the first gene situated 1.1Kb away from the *frCdx4*, followed by a member of the zinc finger family, the *Lnx2* gene located 3.8Kb from the *frCdx4*.

The mouse and human *Cdx4* has been located on X chromosome; synteny seems to be conserved specially in the 3' region. The nearest gene in the 3' end is *Chic1*; a cysteine rich hydrophobic protein, in human it is located at 108.6Kb and in mouse at 26.2Kb from *Cdx4*. The second nearest gene in the human is an X (inactive) specific gene (*XIST*), placed 241.8Kb from *CDX4*; in the mouse is a testis specific gene (*Tsx*) that is positioned at 48.4Kb from its respective *Cdx4* (Figure 3.3).

Synteny appears to be broken in the genes located at the 5' region of the human and mouse *Cdx4*. In human, the immediate gene is a member of the C1 family, the *DMRTC1*, which is located 598.4Kb from *CDX4*, the second nearest gene is peptidyl-prolyl cis-trans isomerase gene (*PIN4*) positioned 1.17Mb from *CDX4*. in the mouse, the nearest 5' gene is *Ppnx* gene placed 64.7Kb from *Cdx4*, the second nearest gene is the riken gene (*Rik*) at 338.9Kb from *Cdx4* (Figure 3.3).

The transcriptional organisation of the *Fugu Cdx4* showed to some extent conservation with the human and mouse organisation, mainly at the 3' flanking region; the *Chic1* gene is conserved in the three species as the nearest 3' gene. However, conservation in the 5' end is different even for the mouse and human *Cdx4*.

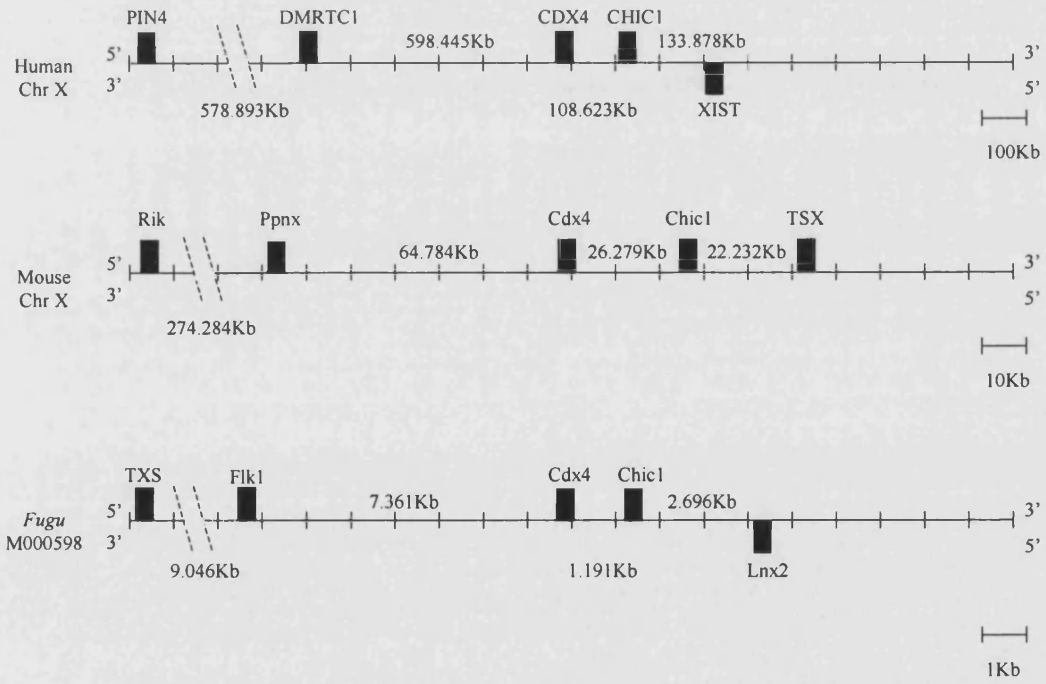


Figure. 3.3. Transcriptional orientation of the *Cdx4* gene in human mouse and *Fugu*.

The *Cdx4* genomic regions from human and mouse were extracted from

Ensembl. The genomic DNA extracted was up to the two nearest neighbouring genes. The *Fugu Cdx4* genomic region (M000598) was extracted from the *Fugu* database. Black boxes indicate the position of each gene through the chromosome, genes located above the line are transcribe 5'→3', genes located under the line are positioned in the reverse DNA strand and are transcribed in opposite orientation. The scale bar for the human is 100Kb, mouse is 10Kb and *Fugu* is 1Kb.

3.3.3 Gene organisation of the *Fugu Cdx* genes

3.3.3.1 Gene structure of *Fugu Cdx1*

The gene structure of the *Fugu Cdx* genes is well maintained with their orthologues in human and mouse. The *frCdx* genes also contain three coding exons and two introns; the introns show similar characteristics to the mammalian *Cdx*, the first intron is much larger than the second intron (Figure 3.4).

In order to obtain the gene structure of the *Fugu Cdx* genes, the intron-exon boundaries were the base to delimit the length of each exon. The intron-exon faces are perfectly conserved between the three species. The first exon, the intron-exon face is +1 (G GG) and for the second exon, the intron-exon face is +0 (CAG).

The human *Cdx1* is over 17.705Kb, contains 3 coding exons, 2 introns and 5' and 3' UTRs. The 5' UTR of the gene is 81bp long and is located upstream of the first exon. This exon extends over 445bp and is followed by the 1st intron, the largest in the gene with 15446pb in length. The second exon is 146bp in length. The 2nd intron, situated between the second and third exons, extends over 560bp in length. The third exon is the final coding exon of the gene and is 207bp in length. After the third exon, the 3' UTR extends over 820bp.

The mouse *Cdx1* shows a very high level of conservation with its human orthologue; the structure is maintained with 3 coding exons, two introns and the 5' and 3' UTRs. The length of the gene is 17.338Kb. The 5'UTR is 79bp in length and is located upstream of the first coding exon which is 445bp in length. The 1st intron positioned between the first and second exon is also the largest of the gene with 15156bp in length. The second coding exon, 146bp in length, is followed by the 2nd

intron 432bp in length; the final coding exon extends over 215bp downstream of the final coding exon and the 3'UTR extends over 865bp.

The *Fugu Cdx1* gene, extending over 2.893Kb, is around 5.8 times smaller than its orthologues. The gene structure of the *frCdx1* is similar to human and mouse, having three coding exons, 2 introns and 5'UTR. The 5'UTR is 81bp in length, the same as the human 5'UTR, and 2bp longer than the mouse 5'UTR. The first exon extends over 398pb, 47bp shorter than the mouse and human ones. The 1st intron is 1436bp, one tenth the size of the human and mouse 1st intron.

The second exon is 146bp in length, the same length as the human and mouse second exons. The 2nd intron is located between the second and third exons, and is 703bp in length; this second intron is 143pb and 271bp larger than the human and mouse 2nd introns respectively. The third coding exon extends over 180bp, 35bp shorter than the mouse third exon and 27bp shorter than the human third exon. The homeodomain is also encoded by the second and third exons as in the case of mouse and human (Figure 3.4).

The 3'UTR of the *Fugu Cdx1* has not been described yet, however, due to the high conservation level that the *frCdx1* shows with its orthologues, the existence of a conserved 3'UTR is probable.

Comparison of the nucleotide sequence shows that the 5'UTR is 32.09% conserved between the three species (26/81pb). Between mouse and human the conservation is 87%, *Fugu* and mouse or *Fugu* and human show 14% identity in the 5'UTR. The first exon is 40.70% conserved with mouse and human (162/398bp). The second exon, which has the same length for the three species, has 70.57% conservation with the mouse and human second exon (103/146pb). The third exon shows 48.33% conservation between species (87/180bp). The conservation of the homeodomain is 69.44% between the three species (125/180bp). To compare the 3'UTR, 861bp downstream of the third coding exon of the *frCdx1* were used for comparison against the human and mouse 3'UTRs. Mouse and human have 66% conservation. The human and *Fugu* show 4% conservation, mouse and *Fugu* have 2% conservation in their 3'UTRs (Appendix section 1A).

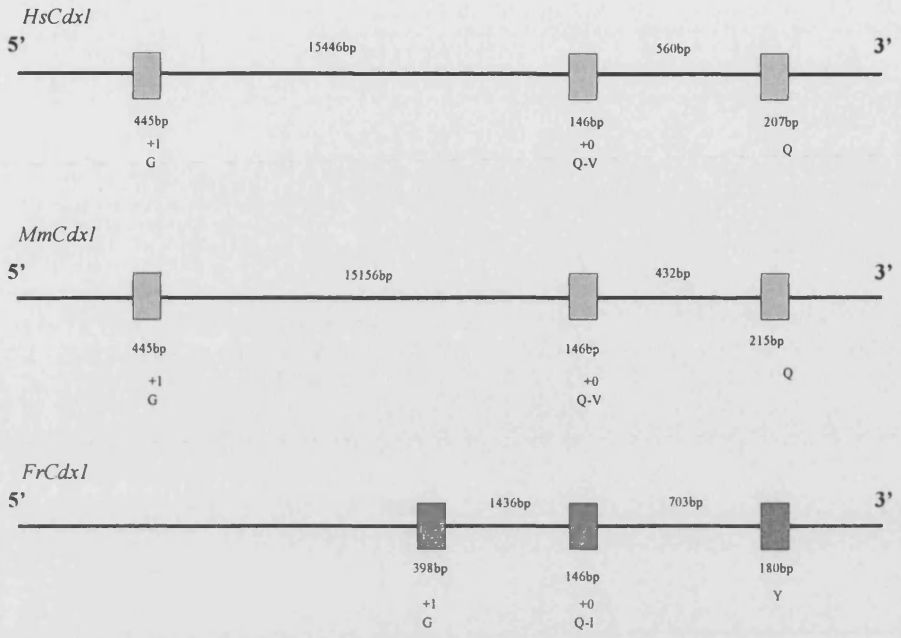


Figure. 3.4. Comparison of the human mouse and *Fugu Cdx1* genes. Filled boxes represent coding exons. The length of each exon and intron is indicated. The splicing acceptor and the intron exon face are shown.

3.3.3.2 Gene structure of *Fugu Cdx2*

The human and mouse *Cdx2* have been also sequenced and characterized. Their structures are well conserved with three coding exon and two introns, the 1st intron being larger than the 2nd one. In the mouse the first coding exon extends over 538bp and 541bp in the human; the 1st intron is 3118bp and 3450bp in mouse and human respectively. The second exon extends over 146bp in both human and mouse; the third exon is 252bp in length in mouse and 255bp in human, the human 2nd intron is 1500bp in length and the mouse 2nd intron is 1087bp.

The *Fugu Cdx2* was the most incomplete of the three *frCdx* genes; the comparative analysis using the mouse and human Cdx2 transcripts and aa. sequences showed the presence of the 2nd and 3rd coding exons. However we were unable to identify the 1st coding exon by a bioinformatic approach. We resorted to the 5' RACE technique to try to identify the first exon of the *frCdx2* gene.

3.3.3.3 Race of the *Fugu Cdx2*

Specific primers for the 2nd exon of the *frCdx2* gene were designed; we used total mRNA extracted from adult *Fugu* gut, the BD SMART™ RACE cDNA amplification kit from BD Biosciences Clontech. Specific primers for the *Fugu Actin* gene were used as internal positive controls. The cDNA was prepared using the modified oligo (dT) primer; the integrity and quality of the cDNA was checked using specific primers for the *Fugu Actin* (Figure 3.5, line 6). The PCR product showed that the cDNA quality was good; a sharp band was observed at 480bp, which is the size expected; no distinctive genomic DNA was observed.

The external positive control was prepared using human placental total RNA. Using the cDNA prepared and specific primers for the transferrin receptor gene, a PCR was performed to produce a 2.6Kb PCR fragment (Figure 3.5, lane 5).

Negative controls were done using in one reaction the specific primers for the gene GSP1 and NGSP1 without the addition of the UMP primer (Figure 3.5, line 3). The second control was prepared using UMP only without specific gene primers (Figure 3.5, line 4); these controls are meant to not generate any PCR product.

The cDNA prepared from the gut mRNA was used to prepare the 5' RACE PCR reaction, after the primary PCR amplification, a smear product was obtained (Figure 3.5, line 1). A second 5' RACE PCR reaction was prepared using as a template a 1/50 dilution of the primary PCR and a nested primer specific for the gene. Two distinctive PCR products were obtained, an 800bp band and an approx. 350bp band. The larger band matched with the sized expected. In the mouse *Cdx2*, the 1st exon extends over 538bp, whereas the 5'UTR is 267bp in length, this results in an 805bp fragment. The larger band was purified and subcloned into pGEM vector and positive clones were sent to sequence.

The sequence obtained was translated into three different frames using the EMBOSS program. The sequencing result showed that the larger PCR product obtained by 5'RACE is effectively the *frCdx2* gene. Unfortunately, the product obtained was a genomic fragment of the gene, which contains the region of the second exon where the GSP1 and NGSP1specific primers were designed and a region that belongs to the first intron of the gene.

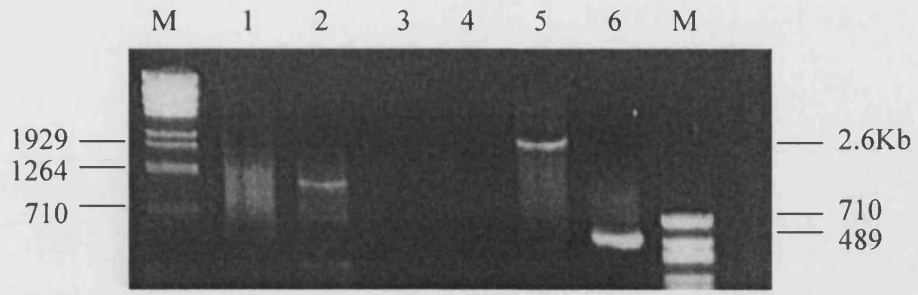


Figure. 3.5. Agarose gel electrophoresis of 5'-RACE product. Total RNA from gut adult *Fugu* was used to prepare the first strand synthesis. Following the first 5'-RACE PCR amplification, which produced a smear like product (line 1). A nested PCR produced an 800bp approx. band (line 2). The negative controls were performed using the GSP1 and NGSP1 (line 3) and UMP only (line 4). External positive control was performed using the cDNA from human placental total RNA and specific primers for the transferrin receptor gene, which produced a 2.6Kb PCR fragment (line 5). The internal positive control was performed using the gut adult *Fugu* cDNA and specific primers for the *Actin* gene to produced a 480bp PCR fragment (line 6). Lanes M indicate the molecular mass markers. The sizes (in bp) of the molecular markers are indicated.

Based on the sequence obtained from the *Fugu* genome database and the comparative analysis using the mouse and human *Cdx2*, the *frCdx2* second coding exon extends over 146bp; the 2nd intron located between the second and last exon is 121bp in length and the third exon is 222bp, 33bp shorter than the human exon.

Using the NIX program, the best open reading frame was located at 1136bp upstream of the second exon, the putative first coding exon, which has the same open reading frame as the mouse and human *Cdx2*. It extends over 67bp, the 1st putative intron extends over 1074bp in length (Figure 3.6).

A comparative analysis of the nucleotide sequence shows that the *frCdx2* second coding exon, which shares the same length as that of the human and mouse exon is 76.02% conserved (111/146bp) with the mouse and human second exon. The third exon is 52.70% conserved (117/222bp) among species. The first putative first exon is 40.29% conserved (27/67bp), using the *Fugu* sequence as base of comparison, although this has still to be verified. The *frCdx2* homeobox is 67.77% conserved (122/180bp) with the mouse and human homeobox (Appendix section 1B).

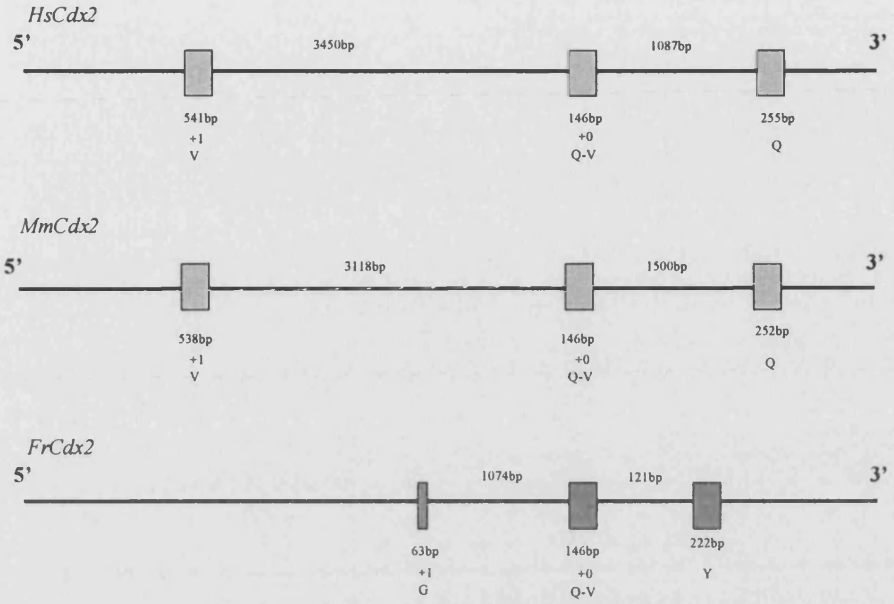


Figure. 3.6. Comparison of the human mouse and *Fugu Cdx2* gene. Filled boxes represent coding exons the length of each exon an intron is indicated. The splicing acceptor and the intron exon face are shown. The *frCdx2* first exon is still to be confirmed

3.3.3.4 Gene structure of *Fugu Cdx4*

The human and mouse *Cdx4* also preserves the characteristic gene structure of the *Caudal* genes. Three coding exons and two introns are located between each exon. The human first coding exon is 496bp and the mouse is 503bp. The 1st intron is the largest intron of the gene, although not as large as the first *Cdx1* intron. The human 1st intron extends over 5.953Kb whereas the mouse 1st intron extends over 5.761Kb. The second exon is 146bp in length for both human and mouse; the human 2nd intron is larger with 7.166Kb in length compare with the mouse, which is 1.279Kb in length. The third exon is 207bp in length in both human and mouse.

The *Fugu Cdx4* gene extends over 2.473Kb in length; 3.26 times shorter that the mouse *Cdx4* and 5.57 times shorter than the human *CDX4*. The gene contains three coding exons and two introns. The 5' and 3' UTRs of the *frCdx4* have not been mapped yet. The first exon is 432bp in length; 64bp and 71bp shorter than the mouse and human respectively. The 1st intron, the largest of the gene is 1.148Kb, however, is

almost 5.3 times shorter than the mouse and human 1st intron. The second exon is 146bp in length, the 2nd intron is located between the second and third exon, extending over 493bp in length. The last coding exon is 207 bp in length. The second and third exons are the same length compared to the human and mouse exons whereas the 2nd intron is almost 2.59 shorter than the mouse and 14.53 times shorter than the human 2nd intron (Figure 3.7).

Analyses of the nucleotide sequence of the *frCdx4* exons of the gene shows that the first exon is 40.87% conserved (177/433bp) compared with mouse and human. The second exon is 60% conserved (106/146bp) among species and the third coding exon is 57.97% conserved (120/207bp) with the mouse and human third exon. The homeobox shows a very high conservation 77.77% (140/180bp); however the hexapeptide domain is only 16.16% conserved (3/18bp) among species (Appendix section 1C).

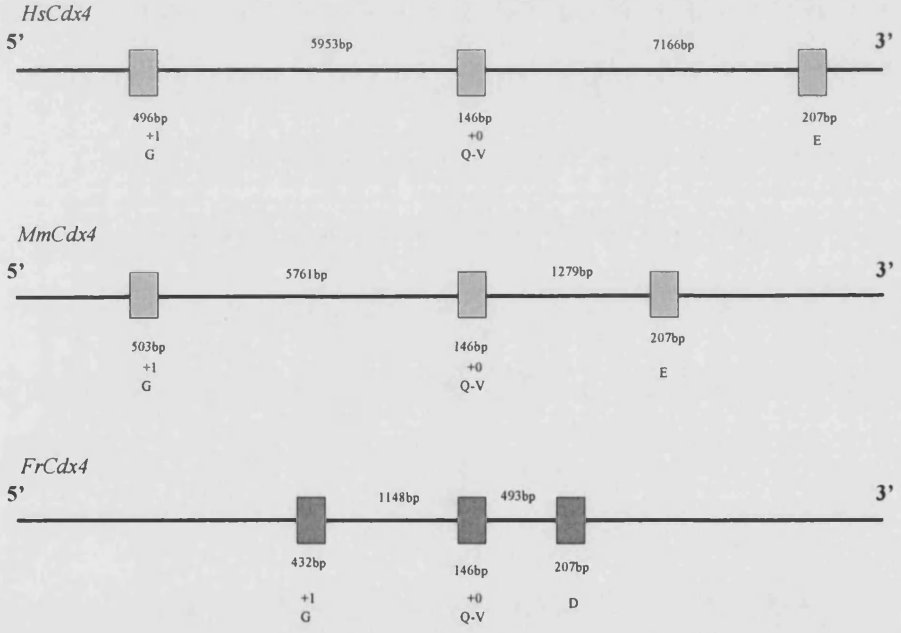


Figure. 3.7. Comparison of the human mouse and *Fugu Cdx1* gene. Filled boxes represent coding exons the length of each exon an intron is indicated. The splicing acceptor and the intron exon face are shown.

3.3.4 Expression analysis of Fugu Cdx by RT-PCR

3.3.4.1 Expression of *Fugu Cdx1*

The second phase of expression of mouse and human *Cdx1* is restricted to embryonic and adult intestinal tissue. To examine the tissue distribution of *Fugu Cdx1*, an RT-PCR was performed using RNA extracted from various *Fugu* adult tissues. The RT-PCR showed that the *Fugu Cdx1* is restricted to the gut adult tissue; the other tested tissues, gonads, gills, kidney, spinal cord, spleen, eye and heart were shown to be negative in expression for the gene (Figure 3.8). Although the RT-PCR performed does not quantitate the amount of gene expression, the amount of product produced by the gut *frCdx1* compared with the amount of product produced by the *Actin* appears to be roughly equal, which may indicate that *frCdx1* is well expressed in the adult gut tissue in *Fugu*.

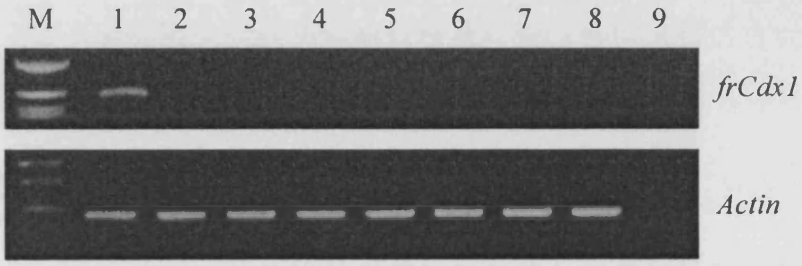


Figure. 3.8. *Fugu rubripes Cdx1* mRNA in adult tissues. The *Fugu Cdx1* mRNA-specific primers amplified a PCR product from total RNA extracted from the gut(1), gonads(2), gill (3), kidney(4), spinal cord (5), spleen(6), eye(7) heart(8)and no template(9). A RT-PCR product corresponding to the constitutively expressed *Actin* gene was used as internal control.

3.3.4.2 Expression of *Fugu Cdx2*

The expression of mouse and human *Cdx2* is restricted to the small intestine and colon in the adult organism. To investigate the expression of *Cdx2* in *Fugu*, an RT-PCR was performed using RNA extracted from different *Fugu* adult tissues. Primers specific for the *Actin* gene were designed to check quantity and quality of the cDNA prepared. The RT-PCR showed that the *frCdx2* is specifically expressed in the adult gut tissue in *Fugu*; gonads, gill, kidney, spinal cord, spleen, eye and heart were

negative for the *frCdx2* expression (Figure 3.9). The RT-PCR also showed that the amount of product produced for the *frCdx2* is less than the amount produced for the *Actin* gene. This may be an indication that the *frCdx2* is not so highly expressed in the *Fugu* gut.

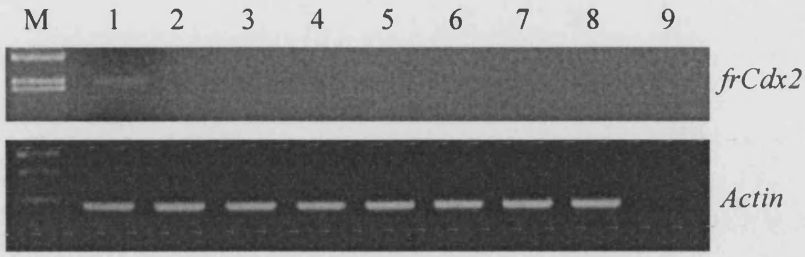


Figure. 3.9. *Fugu rubripes Cdx2* mRNA in adult tissues. The *Fugu Cdx2* mRNA-specific primers amplified a PCR product from total RNA extracted from the gut(1), gonads(2), gill (3), kidney(4), spinal cord (5), spleen(6), eye(7) heart(8) and no template(9). A RT-PCR product corresponding to the constitutively expressed *Actin* gene was used as internal control.

3.3.4.3 Expression of *Fugu Cdx4*

The third *Caudal* gene, the *Cdx4*, is highly active during the early stages of development. However, no expression has been detected after 10.5dpc in mouse development. To investigate *frCdx4* expression in *Fugu*, an RT-PCR was carried out; RNA from different adult *Fugu* tissues was extracted and analysed for the expression of the *frCdx4*. The result showed that *frCdx4* is not expressed in any of the *Fugu* adult tissues tested (gut, gonads, kidney, spinal cord, spleen eye and heart). The product obtained for the *Actin* gene showed that the cDNAs prepared were of good quality (Figure 3.10). This result indicates that the *frCdx4* is not expressed in the adult organism as with the mouse and human *Cdx4*.

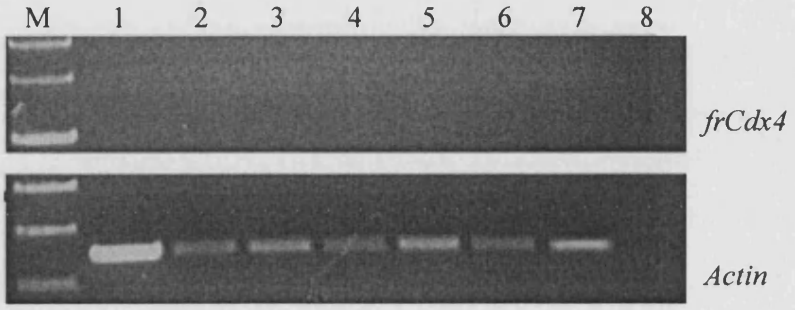


Figure. 3.10. *Fugu rubripes Cdx4* mRNA in adult tissues. The *Fugu Cdx4* mRNA-specific primers amplified a PCR product from total RNA extracted from the gut(1), gonads (2), kidney(3), spinal cord(4), spleen(5), eye(6), heart(7) and no template(8). A RT-PCR product corresponding to the constitutively expressed *Actin* gene was used as internal control.

3.3.5 Analysis of the Fugu Cdx protein sequence

3.3.5.1 Amino Acid sequence of *Fugu* Cdx1

The human CDX1 protein is 265aa in length. The first exon codes for 152aa, the second and third ones codes for 51 and 62aa respectively. The mouse Cdx1 protein is 268aa, three amino acids longer than the human CDX1; the first and second coding exons are conserved in length with the human. The difference is in the third coding exon, which codes for 65aa.

The *frCdx1* protein sequence was predicted using the genomic sequence of the gene and the predicted transcript of the gene. It extends over 225aa in length; it is 40 and 43aa shorter than the human and mouse Cdx1 respectively. The first exon encodes for 120aa, 32aa shorter than the mouse and human first coding exons. The second coding exon extends over 51aa, which is the same length for the mouse and human. The third coding exon is 54aa in length, 8aa shorter than the human and 11aa shorter than the mouse third coding exons.

The protein sequence of the *Fugu Cdx1* shows an overall conservation of 74.22% (167/225aa) with the human and mouse *Cdx1*, considering the identical aa. and the conserved substitutions. The first coding exon is 65.83% conserved (79/120aa). The second exon shows 96.07% conservation (49/51aa) and the third coding exon is 72.22% (39/54aa) conserved between the three species, when using the *Fugu* aa. sequence as a basis for comparison.

Analysis of the characteristic motifs of the Cdx1 reflects a good level of conservation of the Cdx1 across species. The homeodomain, the main domain of the protein, which represents almost ¼ of the complete protein, shows 96.66% (58/60aa) of conservation. The hexapeptide domain (Hex domain); which is characteristic of the Homebox proteins is 66.66% (4/6aa.) conserved in *Fugu*; it is located 16 residues

upstream of the homeobox, exactly in the same position as the human and mouse Hex domain. Although the conserved residues present in the *Fugu* Hex domain are not entirely identical, they are amino acids with similar characteristics to the mouse and human Hex domain (Figure 3.11).

Specific domains show a high degree of conservation; the N terminal arm, which marks the start of the homeodomain, is 100% (8/8aa) conserved; the phosphorylation sites SGK, TIRRK and TER are 100% conserved; in the second phosphorylation site in *Fugu* the isoleucine is substituted by a methionine (I→M), another conserved substitution is present in the TER domain where the threonine changed to serine (T→S). The turn domain (3/3aa), and the phosphatase-binding site (4/4aa) are 100% conserved between the three species. The *Fugu* acetylation site, although with a 100% (6/6aa) conservation with the human and mouse, shows two substitutions in its sequence, the valine changed to methionine (V→M) and the second lysine to arginine (K→R), see figure 3.11.

3.3.5.2 Amino Acid sequence of *Fugu* Cdx2

Using the mouse and human Cdx2 amino acid sequence, the putative *Fugu* Cdx2 aa sequence was characterized by comparison. Firstly, human CDX2 is a 313aa sequence, 2aa longer than mouse Cdx2. The difference in those 2aa, arises because the human first and third exons are one aa longer than the mouse ones with 184 and 78aa respectively. The second exon is the same length in mouse and human with 51aa.

The *Fugu* Cdx2 protein was predicted to extend over 146aa, almost two times shorter than the human and mouse Cdx2. However, it has 78.76% conservation (115/146aa) with mouse and human. The prediction for the first exon translates in to 27aa, almost 7 times smaller than the human first exon. The second *Fugu* exon is 51aa in length, the same length than the mouse and human second exon. The third exon is 68aa in length, 10aa shorter than the human and 9aa shorter than mouse third exon (Figure 3.12).

The first *Fugu* exon is 77.77% (21/27aa.) conserved with mouse and human, using the *Fugu* protein as a base sequence for comparison. However, as mentioned earlier this may not be the first exon or just a part of it. The second exon is 100% conserved (51/51aa.) and the third exon shows 63.23% conservation (43/68aa.) with the mouse and human third exon. The frCdx2 homeodomain shows a very high level of conservation, it is 100% conserved (60/60aa.) compared with mouse and human.

The specific motifs in the Cdx2 protein are also well conserved in the frCdx2. The N- terminal region is 100% conserved as well as the turn domain and the phosphatase-binding site. The putative phosphorylation sites of the protein, TIRRK and SER are also 100% conserved among species. The acetylation site is 100% conserved with the human; the mouse acetylation domain is one aa shorter. The glutamine region after the acetylation site seems to be reduced to only two aa in *Fugu* (Figure 3.12).

3.3.5.3 Amino Acid sequence of *Fugu* Cdx4

The Cdx4 aa sequence has been also fully characterized for mouse and human. Human CDX4, as for the human CDX2, is 2aa longer than the mouse Cdx4. The first human exon is 171aa, whereas the mouse first exon is only 169aa. The second and third exons are conserved in length between mouse and human with 51 and 62aa respectively.

Fugu Cdx4, the longest of the *Fugu* Cdx, extends over 267aa. It is 17aa shorter than the human and 15aa shorter than the mouse Cdx4. The first exon is 148aa in length, 23 and 21aa shorter than the mouse and human respectively. The second exon as all the Cdx genes is 51aa in length; the final exon is 68aa, interestingly 6aa. longer than the mouse and human third exon. Taking together the identical amino acids and the conserved substitutions, the frCdx4 has 76.77% conservation with the mouse and human. The first *Fugu* exon is 72.97% conserved (108/148aa), the second exons shows 98.03% conservation (50/51aa) and the third exon 69.11% conservation (47/68aa) among species (Figure 3.13).

The homeodomain is encoded by the second and third exons, it is 98.33% conserved 59/60aa among species. The Hex motif is 12 residues upstream of the homeodomain. The Hex motif appears to be reduced to 5 residues in *Fugu*; it is 66.66% conserved (4/6aa). Conserved substitutions have taken place in the *Fugu* Hex motif. The tyrosine has been substituted to phenylalanine (Y→F), the alanine to glutamine (A→Q) and the arginine to aspargine (R→N).

Specific domains in the protein also show high conservation when compare to the human and mouse Cdx4; the phosphorylation sites TGK, TIRRK AND SER are 100% conserved, the turn domain and the phospahtase binding site are also 100% conserved. The acetylation site shows 100% conservation with only one conserved substitution, the second methionine has changed to leucine (M→L), the rest of the motif remains adentical (Figure 3.13).

Figure. 3.11. Clustal W alignment of Cdx1 amino acid sequences

Alignment of motifs are shown: Phosphorylation site (purple box), Turn domain (green box), Phosphatase binding site (yellow box), N terminal arm (red box), hexapeptide domain (grey box), KIKK acetylation site (pink box) and Glutamine region (blue Q). The amino acid sequences were taken from Ensembl: *Cdx1* human ENSP00000231656, *Cdx1* mouse ENSMUSP00000025521

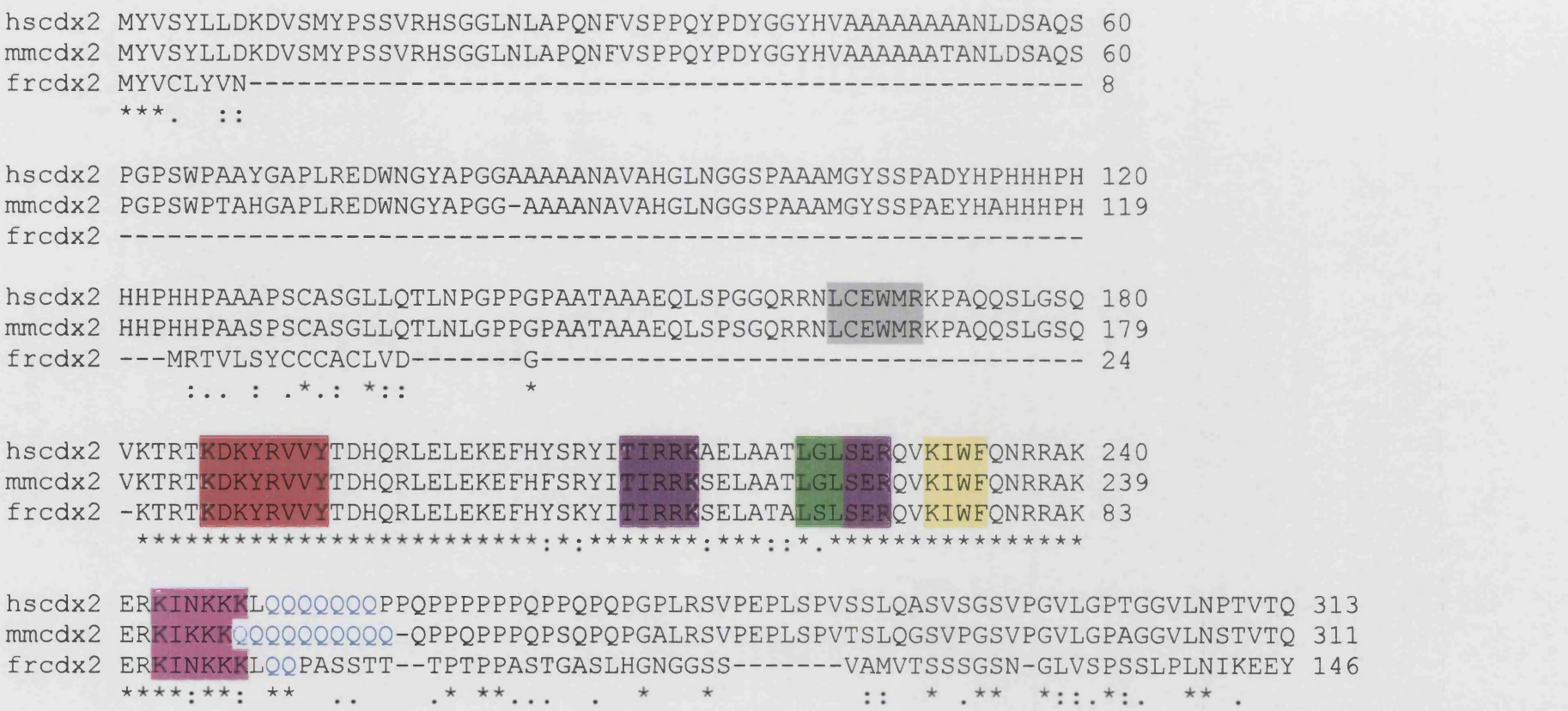


Figure. 3.12. Clustal W alignment of Cdx2 amino acid sequences

Alignment of motifs are shown: Phosphorylation site (purple box), Turn domain (green box), Phosphatase binding site (yellow box), N terminal arm (red box), hexapeptide domain (grey box), KIKK acetylation site (pink box) and Glutamine region (blue Q). The amino acid sequences were taken from Ensembl: *Cdx2* human ENSP00000298359, *Cdx2* mouse ENSMUSP00000031650

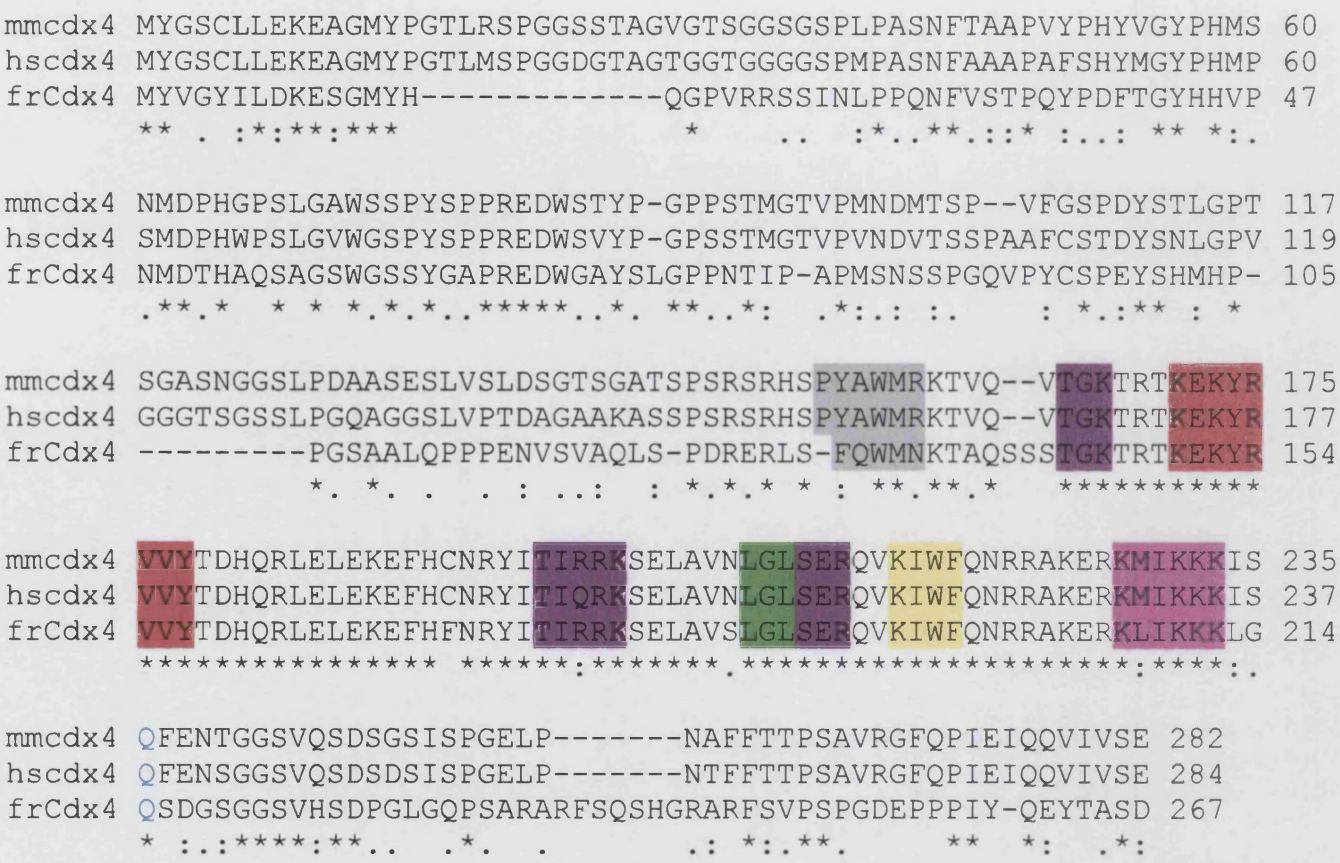


Figure. 3.13. Clustal W alignment of Cdx4 amino acid sequences
Alignment of motifs are shown: Phosphorylation site (purple box), Turn domain (green box), Phosphatase binding site (yellow box), N terminal arm (red box), hexapeptide domain (grey box), KIKK acetylation site (pink box) and Glutamine region (blue Q). The amino acid sequences were taken from Ensembl: *Cdx4* human ENSP00000253572, *Cdx4* mouse ENSMUSP00000033689

3.3.6 Discussion

The *Fugu* genome has been primarily used for the detection and characterisation of genes in higher organisms, and recently for the identification of conserved non-coding regions involved in the regulation of genes. In my search for the regulatory elements involved in the regulation of *mCdx1*, we used the pufferfish as a comparative model. The comparative analyses performed using bioinformatic approaches showed that the three *Caudal* genes are present in *Fugu*; linkage, gene structure and sequence were well conserved when compared to the mouse and human *Cdx* genes. Amino acid sequence comparison and expression of the genes were also found to be conserved across species

The *Fugu Cdx* genes were identified and categorised based on their linkage conservation. The *frCdx1* was identified by its linkage with the *Pdgfrβ* and *Csf1r* genes, located at the 5' end of the gene. Early reports show a conserved linkage between these *Fugu* and human tyrosine kinase receptors, *PDGFRβ* and *CSF1L* (How *et al.* 1996). However, this work did not describe the linkage between *Pdgfrβ* and *Cdx1* and their conservation with human. Here we show the linkage conservation between the *Fugu Cdx1* and *Pdgfrβ*. This conservation is also maintained in mouse and human as shown by the physical maps from the Ensembl database.

The *Fugu Cdx2* was identified mainly by its linkage with the insulin promoter factor (*Ipf1*). Studies performed in Amphioxus showed that *Cdx2* belongs to the Parahox genes, a cluster of genes originated by duplication that resemble the Hox genes and are expressed in an anterior posterior spatial co-linearity. The Parahox cluster is formed by *Gsh1*, a gene expressed in the forebrain and midbrain; *Ipf1* (also known as *Pdx1*), expressed in the pancreatic β cells and in duodenum; and the *Cdx2* gene, expressed in the proximal colon with a gradual expression in the ileum and jejunum (Brooke *et al.* 1998). The *frCdx2* seems to have the 3' end region conserved with the insulin precursor (*IP*) gene, which appears to be the *Fugu* homologue to the *Ipf1* gene. However, the *IP* gene is placed as the second proximal gene to the *frCdx2*; the ring finger protein 26 (*RFP*) is the most proximal 3' gene to the *frCdx2*.

For the *Fugu Cdx4*, transcriptional organisation was conserved principally at the 3' flanking region. The *Chic1* gene is conserved in mouse, human and *Fugu* as the

nearest 3' gene. Synteny seems to be lost in the 5' end, not only between *Fugu* and mouse or human, but also between human and mouse.

The gene structure of the *frCdx* genes seems to be well conserved. Three coding exons (except for *frCdx2*) and two introns placed between each exons is the general structure of the *Caudal* genes. Nevertheless, the first exon was slightly shorter, at least in the case of the *frCdx1* and *frCdx4*; this was reflected in the low nucleotide conservation of the exon. This is not the case for the second and third exons, especially for the second exon, where length and nucleotide sequence are better conserved. The length of the introns also showed the characteristic size reduction of the *Fugu* genes, although the 1st intron of the *Cdx* remains longer than the 2nd intron. Like the human and mouse homeodomains, the *Fugu* homeodomains are coded by the second and third exons; in the three *Fugu* genes, they extend over 180bp and show a high degree of conservation with the mouse and human homologues.

Even though we were not able to obtain the first coding exon of the *frCdx2*, the 5' RACE result showed that the designed GSP1 and NGSP1 primers were very specific for the *frCdx2* gene. Sequencing of the product showed that 120bp belong to the second exon, and that the 1st intron extends at least over 780bp. One of the key steps in the identification of the 5'end of the genes is to avoid and remove any genomic contamination. This is achieved by treating the RNA with DNAseI. Despite the fact that the samples were treated with DNAseI, a complete degradation of the genomic material was not achieved; this could have resulted in the use of genomic DNA as a template, instead of the cDNA.

The expression analyses of the *Fugu Cdx* genes showed that, as in human and mouse, the expression of these genes is restricted to the gut tissue in the adult pufferfish, in the case of *frCdx1* and *frCdx2*. The absence of *frCdx4* expression in the *Fugu* adult tissue also reflects the conservation of function with the mouse *Cdx4*, where no expression has been detected after 10.5dpc. Further experiments using *Fugu* RNA from different embryonic stages will indicate if the expression of the *frCdx* genes is also present in the early stages of development and if it is comparable to the expression of other organisms.

Based on the coding sequence obtained, the amino acid sequence of the *Fugu* Cdx proteins showed a high degree of conservation among species. Especially, the specific domains of the protein are 100% conserved and the hexapeptide domain, which is a characteristic feature of all Cdx proteins, is conserved between each *Fugu* Cdx and its homologue. Furthermore, the Hex domain of the zebrafish *Zcad1* (SYQWMS), which has been described as the *Cdx1* homologue in zebrafish, seems more similar to the *Fugu Cdx4* (FQWMN). The frCdx4 Hex is 83.33% conserved (5/6aa.) with the Zcad1 Hex domain.

The *frCdx1* and *frCdx4*, which showed the same gene structure, seem to have a very short first exon that is translated into a shorter peptide, compared to the mouse and human Cdx1 and Cdx4 proteins. The frCdx1 and frCdx4 proteins appear to have lost the regions rich in proline residues, which are codified by the first exon of the gene.

Even though the frCdx2 shows a lack of the first peptide region, the start of the N- terminal region is conserved with the human and mouse Cdx2. The frCdx2 homedomain is the most conserved of the *Fugu* Cdx with 100% sequence identity with the human and mouse Cdx2. The glutamine residues located after the acetylation domain, although reduced to two residues, are also conserved in the frCdx2.

Using comparative genomics and a set of bioinformatic tools, we have identified and characterized almost completely the *Fugu Cdx* genes from the *Fugu* genomics project database. We found that the three *Cdx* genes are present in the *Fugu* genome, that the physical linkage of the *Cdx* and neighbouring genes are conserved in *Fugu*, mouse and human, and that the structure of the *frCdx* genes and proteins as well as their expression are also conserved with human and mouse. A functional characterization of the *Fugu Cdx* genes will corroborate the role of these genes in the *in vivo* organism.

Chapter Four

Identification of the Cis-regulatory elements and expression analyses of the mouse and *Fugu Cdx1*

4.1 Introduction

4.1.1 Early zebrafish development

The development of the zebrafish embryo, like other cyprinids, involves the main processes present in mammalian development. The zebrafish undergoes seven main stages through its development- the zygote, cleavage, blastula, gastrula, segmentation, pharyngula and hatching. Contrary to mammalian development, zebrafish development occurs very quickly after the egg is fertilized; in 72hrs the embryo has almost completed its development. The new larva is able to swim, seek for food and adopt avoiding behaviours.

The zygote and cleavage periods occur in the first 2 hours after fertilization, where cell division and arrangement of blastomeres take place. The cleavage does not occur in the yolk region, the blastomeres are located above the yolk. The final cleavage results in the blastoderm, with a single outer layer of flattened cells and a inner layer of rounded cells (Wolpert *et al.* 2002) During the 2 to $4^{3/4}$ hpf, the embryo is in the blastula period. The blastoderm expands to the vegetal pole to cover the yolk in a process known as epiboly. Gastrulation begins by $5^{1/2}$hpf, during which the future endodermal and mesodermal cells involute or move interiorly at the margin of the blastoderm. These cells migrate towards the future dorsal side. The embryo begins to elongate in an anterior posterior direction (Wolpert *et al.* 2002).

During the 80% epiboly, which is between 8 and 9hpf, the dorsal side of the embryo becomes distinctly thicker, the brain rudiment begins to thicken and the notochord rudiment becomes distinguishable from the segmental plate.

By the 14 somite stage, the embryo is 0.9mm. The otic placode, brain neuromeres and trunk somites become visible, and the yolk extension is starting to form and elongate; the pronephric duct is not visible at this stage.

The hatching period of the zebrafish embryo occurs at 48hpf; by now the embryo is 3.1mm in length, the yolk extension begins to elongate, the melanophores begin to appear in the lateral stripe and the iridophores are present in the retina. Blood circulation is visible in the aortic arches and segmental vessels. The liver and the swim bladder appear along with the gut tract. The gut tract locates posterior to the pharynx in the region of the anterior yolk ball. At 39hpf, the hindgut endoderm is

recognizable as a solid plug of midline cells in the anal region more than a hollow tube, which looks thicker than the adjacent pronephric ducts (Kimmel *et al.* 1995).

By day 3 the hatched larva has completed most of its morphogenesis, and it continues to grow rapidly. The mouth continues to open and elongate interiorly. By day 4, the swim bladder inflates. The gut tube moves more ventrally and the yolk extension nearly empties. The larva swims more actively, moves the jaws, opercular flaps, pectoral fins and eyes (Kimmel *et al.* 1995).

4.1.2 Intestinal development of the zebrafish

The main studies of the zebrafish intestinal formation come from Wallace and Pack (2005) and Wallace et al. (2005). One of the main characteristics of the zebrafish is the absence of a stomach. The adult zebrafish intestine is a folded tube that occupies most of the abdominal cavity. An anterior-posterior axis is also distinguished, and can be subdivided into three segments: anterior, mid and posterior intestine. The anterior intestine, usually referred as the intestinal bulb, has a wider diameter than the posterior intestine and may function as a reservoir.

During zebrafish intestinal development, two stages can be distinguished. In the first stage, the intestine expands rapidly with maturation of polarized epithelial cells along with smooth muscle and enteric nerve cells. In the second stage, smooth muscle, enteric nervous and epithelial cells differentiate; morphogenesis of the folded epithelium also occurs.

In terms of intestinal morphology, the anterior zebrafish intestine has three of the four cell types present in the mammalian intestine; the columnar shape absorptive enterocytes are the most numerous, present in the anterior and mid intestine. These are followed by the goblet cells, which are present in all the intestinal regions. The enteroendocrine cells are present only in the anterior region of the intestine. The paneth cells have not been identified in any part of the zebrafish intestine (Wallace *et al.* 2005).

The zebrafish intestine is joined to the anterior digestive tract by a short muscular oesophagus. Digestive enzymes such as the intestinal fatty acid protein (Her *et al.* 2004), are mainly expressed in the anterior region of the intestine. The mid intestine lacks enteroendocrine cells; nutrient absorption is possible in this region due

to the presence of enterocytes and the expression of solute transporters and digestive enzymes (Wallace *et al.* 2005). The posterior region of the mid intestine is mainly composed of specialized enterocytes that may function in mucosal immunity; this region resembles the mammalian ileum. The posterior intestine starts just after the posterior mid intestine; it is a short region where the epithelial folds are short and longitudinally arrayed. Absorptive enterocytes are absent in this region, which makes it similar to the mammalian colon (Wallace and Pack 2003; Wallace *et al.* 2005).

Histological studies show that the adult zebrafish intestine is arranged in concentric layers with a lesser degree of complexity than the mammalian intestine. The zebrafish lacks the equivalent to the muscularis mucosa, a smooth muscle layer present in mammals. The enteric nervous system is less complex than in mammals. However, the blood vessels and the surrounding smooth muscle layers are almost identical to those in mammals (Wallace and Pack 2003).

Instead of the characteristic finger- like villi structure, the zebrafish intestinal epithelium is organized in broad irregular folds. The proliferative cells are located at the base of these folds; the crypt structures are not present in the zebrafish intestine, which resembles the mammalian embryonic intestine (Korinek *et al.* 1998). At the base of the epithelial folds there is a mix of differentiated cells and undifferentiated cells. However, it is still unknown if the undifferentiated cells contain multipotent characteristics.

4.1.3 Signalling pathways and factors involved in the posterior zebrafish development

Studies of disruption, overexpression and inhibition of genes in *Xenopus*, mouse and zebrafish have brought to our knowledge the mechanisms and genes involved in the formation of the posterior end of the embryo. As in higher vertebrates, most of the signalling pathways and factors involved in the early processes of the posterior body are present in the zebrafish. Zebrafish tail development is established at the end of gastrulation, through a movement of cells from the ventral and dorsal margins to the vegetal pole (Kanki and Ho 1997).

The first findings about the interaction between FGF, *Cdx* and *Hox* genes and their importance in the formation of the posterior region of the embryo were in experiments performed in *Xenopus* (Pownall *et al.* 1996). Inhibition of FGF in *Xenopus* and zebrafish disrupts the formation of the posterior body; the *no tail* (*ntl*)

and *spadetail* (*spt*) genes which interact with the FGF pathway participate in the formation of the tail in zebrafish (Amaya *et al.* 1993; Griffin *et al.* 1998).

The non-canonical Wnt pathway also participates in the extension movements during gastrulation in zebrafish and *Xenopus* (Veeman *et al.* 2003). The Wnt8 and Wnt3a signalling pathways participate in patterning during embryogenesis. Wnt8 is essential for mesoderm formation during zebrafish gastrulation (Kelly *et al.* 1995). Inhibition of either Wnt8 or the *Dickkopf1* (*Dkk1*) gene reduces the formation of the tail (Lekven *et al.* 2001). Loss of Wnt3a provokes reduction or complete absence of posterior structure and somites in mice (Ikeya and Takada 2001). The Wnt3a null or hypomorphic mutant (vestigial) mice show a reduction in *Cdx1* expression without affecting *Cdx2* or *Cdx4* expression (Ikeya and Takada 2001; Prinos *et al.* 2001). The Wnt3a and Wnt8, and *Cdx1/Cdx4* morphant embryos also fail to promote somitogenesis during mid gastrulation.

The inhibition of *Xcad3*, which is the *Xenopus* orthologue of *Cdx4*, inhibits formation of the posterior body in the frog embryo (Isaacs *et al.* 1998); mutations in the *Cdx4/kugelig* genes reduce the formation of the posterior body in zebrafish (Davidson *et al.* 2003). Although *Cdx1* and *Cdx4* mediate the Wnt-dependent axis formation, *Cdx1* and *Cdx4* expression is dependent on FGF signalling. FGF signalling is required for *Hoxa9a* expression via *Cdx1*.

4.1.4 Cis- regulatory elements and transcription factors

The regulation and expression of genes is a well-regulated and complex process. The instruction for regulation and control of this process is contained in the cis- regulatory modules, which direct the binding of specific transcription factors or signalling molecules to switch on or off the expression of specific genes. The cis-regulatory elements or modules are non-coding genomic DNA located either immediately upstream of the transcription start site of the genes or a few kilobases upstream of the genes; these elements can also be located in the intronic or downstream region of the genes that they control. A cis- regulatory element is constituted of multiple binding sites for transcription factors; these binding sites can be for specific transcription factors or for ubiquitous transcription factors that may be involved in DNA looping, or required to build the basal transcription complex (Arnone *et al.* 1997). Each cis-regulatory element contains several binding sites for

different transcription factors; however, two or more transcription factors must bind to the cis-regulatory sequence to activate or repress the gene (Davidson *et al.* 2002).

Seipel et al. (1992) showed that the transcriptional function of the binding domain is dependent of the binding site location; if the binding site is placed in a proximal or distal position relative to the transcription start site, the transcription factor may be active or not, or may show different transcriptional levels. For instance, p65, NKκB or TFE3, which contain a negative charge domain, or the ITF factors, which contain a serine/threonine domain, are activators in distal and proximal promoters. On the other hand, AP2 and CTF/NF1 (proline domain) and Oct-1, Oct-2 and Sp1 (glutamine domain) factors show mainly proximal promoter activity.

In the case of Cdx2, studies showed that the protein can bind consensus AT motifs in either proximal or distal regions (Suh *et al.* 1994; Troelsen *et al.* 1997; Colnot *et al.* 1998). The Cdx1 factor also binds to the same consensus site as Cdx2, however, this binding might be restricted by other mechanisms and not only the binding site (Moucadel *et al.* 2001). Due to the similar binding site shared by Cdx1 and Cdx2, different factors may be implicated in the interaction with Cdx to discriminate their specific targets (Moucadel *et al.* 2002).

4.1.5 Use of zebrafish in developmental biology

A strategy to test the functionality of the conserved non- coding sequences is the use of a reporter gene expression system in an *in vivo* model. As proposed by Rothenberg (2001), in the search for regulatory regions, two main conditions have to be satisfied: first, the assertion that the essential regulatory sequences are actually present in the DNA to be tested, and second, the use of an expression system that allows us to follow the expression of the investigated genomic sequence.

Both conditions can be difficult to achieve because regulatory elements can be placed upstream or downstream or even in the intronic regions of the gene (Arnone *et al.* 1997), and also because mammalian models designed to look for genes expressed at different developmental stages are costly and slow.

Barton et al. (2001) suggested that by focusing on teleost fishes as sources for genes and assay systems, both of the above conditions can be reached. The zebrafish can provide a valuable model system for the study of gene regulation and development genetics.

To search for the regulatory sequences that control the expression of the *Cdx1* gene, we returned to the use of comparative genomics. This chapter describes the use of *Fugu rubripes* as a model organism to search for the elements involved in the regulation of the *Cdx1* gene in mouse and *Fugu*. Putative cis-regulatory regions from both species were assayed in *in vitro* and *in vivo* systems to identify the transcription factors involved in the regulation of *Cdx1*.

4.2 Material and methods

4.2.1 Sequencing of the Mouse *Cdx1* upstream non- coding region

The sequencing of the *Cdx1* upstream regulatory region was carried out in two sections. The first 5.7Kb, which corresponds to the intron/ exon junctions of the gene, was cloned into BSKS2+ vector using the *Hind III* sites (Figure 4.1). The sequence was provided by Dr Subramanian (unpublished data). For the distal region, a 5.2Kb fragment of the Cdx1 upstream regulatory region was cloned into BSKS2+ vector using the *Hind III* sites (Appendix section 2B). The sequencing was performed from the 5' end of the fragment. The primers were designed according to the new sequence obtained. The primers used were: Cdx1 T7 (5'-aaggaaaagagtggctgat-3'), Cdx1 TT7 (5'-atgtagggcaatagtt-3'), 3T7 Cdx1 (5'-ttctcttccactttc-3'), 4T7 Cdx1 (5'-gtcatagaggagc-3'), 5T7 Cdx1 (5'- aagacacagcgtaga-3') and 6T7 Cdx1 (5'-acagcaggtaagtcag-3'). The primer positions are shown in figure 4.1. The complete sequence of the distal promoter region is shown in the appendix section 2A.

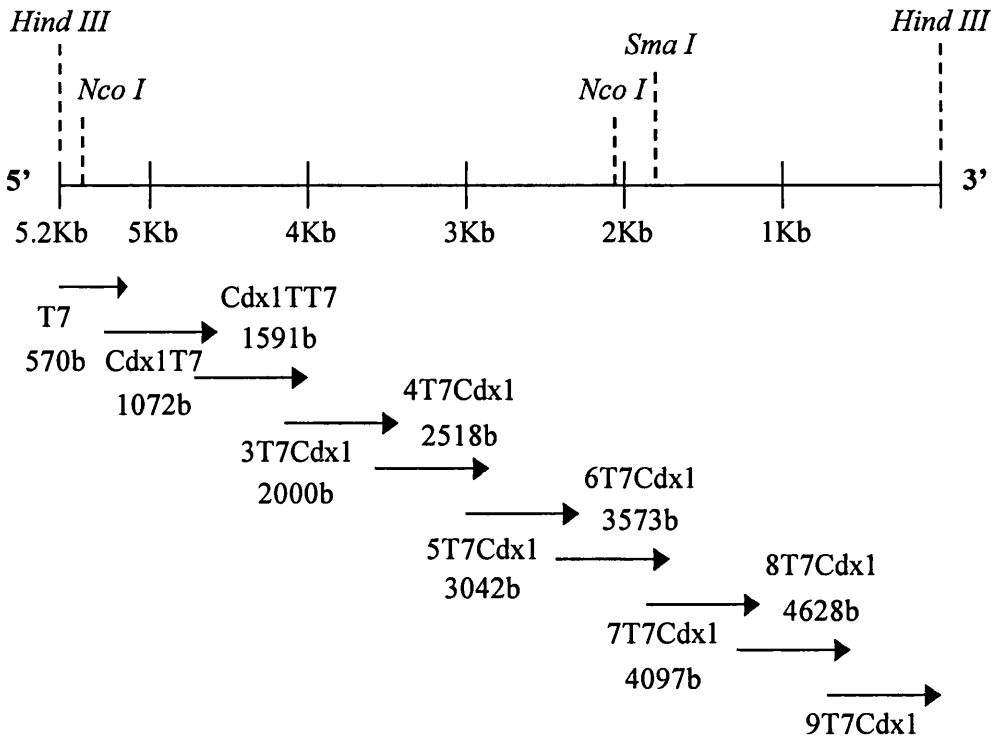


Figure. 4.1. Sequencing of the distal *Cdx1* upstream region. The 5.2Kb fragment cloned into BSKS2+ was sequenced from the 5'end of the insert. The black arrows show the position where the primers were designed and the name of the primer.

4.2.2 Cloning of the upstream sequence of the *Fugu Cdx1*

PCR primers were designed based on the upstream sequence of the *Fugu Cdx1* gene. The oligonucleotide primers synthesized are: 4.0Kb FW frCdx1 (5'-ggg gct gga atg tgt tat ctc taa-3') to produce a 4.0Kb PCR product, 2.0Kb FW frCdx1 (5'-aaa gcc ctg cta gca gta gtt agc-3') to produce a 2.0Kb PCR product and 1.0Kb FW frCdx1 (5'-tca ttg tga ctg atg tat cgg tgc-3') to generate a 1.0Kb PCR fagment. An *EcoRI* site was added to these three forward primers. To the reverse primer, RV frCdx1 (5'-ttc ctg cgt cct ggt gcg taa aat-3') an *XhoI* site was added. The Bac: b238N08 from the Fugu genomic Bac library (Elgar *et al.* 1999) was used as a template.

PCR was performed in a Perkin Elmer thermo- cycler. The reaction mix was as follows: 5µl buffer (10X), 3µl $MgCl_2$ (50mM), 1µl dNTPs (10mM), primers 1µl each (10µM), 1µl DNA and 0.5µl ROCHE™ Taq polymerase (1.5 U/µl) and 37.5µl of H_2O. Amplification was achieved with an initial denaturation at 94°C for 2mins, followed by 35 cycles of 20sec annealing at 60°C, 2min extending at 72°C (2min for

the 1 and 2Kb fragments and 4mins for the 4Kb fragment) and 1min denaturing at 95^{0}C, then a final 5min extension at 72^{0}C. The products of the reaction were checked by running a 2µl aliquot on a 1% agarose 0.5X TBE gel.

4.2.3 Transgene constructs of the *Fugu Cdx1* gene

All the PCR were run in a 1X TAE agarose gel and the band excised and electroeluted. This was then dissolved in 85µl of H_2O, 10µl 10X buffer and 5µl restriction enzyme (5U) and incubated at 37^{0}C for 2hrs. The digested DNA was then phenol/chloroform extracted and precipitated with ethanol. The fragment was then ligated into the *EcoRI/XhoI* site of BlueScript. Constructs were sequenced using the T7 and T3 primers.

The pCS2 CMV-GFP and pCS2 CMV-*LacZ* (Muller *et al.* 2002) were modified by excising the CMV promoter using the *SalI/HindIII* sites. Sites were blunted by T4 polymerase (ROCHE) and the vector was religated.

Construct 4Kb *frCdx1* GFP was generated by subcloning the *EcoRI/XhoI* fragment in BlueScript into the *EcoRI/XhoI* site of the pCS2 GFP^{-CMV}, Construct 2Kb *frCdx*1 GFP was generated by subcloning the *EcoRI/XhoI* fragment in BlueScript into the *EcoRI/XhoI* site of the pCS2 GFP^{-CMV} and Construct 1Kb *frCdx1* GFP was generated by subcloning the *EcoRI/XhoI* fragment in BlueScript into the *EcoRI/XhoI* site of the pCS2 GFP^{-CMV}. Alternatively, the 1Kb and 2Kb fragments were subcloned using the *BamHI/XhoI* into the pCS2 CMV-*LacZ* (Appendix section 2C).

4.2.4 Transgene constructs of the mouse *Cdx1* gene

A 8.9Kb fragment that corresponds to the upstream sequence of the mouse *Cdx1* was cloned into pBSKS+2 (Stratagene) in to separate fragments, a 5.7Kb fragment that harbours the first –3.750Kb upstream sequence of the gene plus the first coding exon. The second fragment cloned into BSKS+2 was a 5.2Kb fragment that contains the –3.751Kb to –8.951Kb upstream sequence of the gene (Subramanian, unpublished data).

Reporter constructs were generated using the pTk*LacZ* vector that contains the herpes simplex thymidine kinase (Tk) minimal promoter, the bacterial *LacZ* gene and the SV40 polyadenylation site (obtained from V. Subramanian). The 5.2Kb mCdx1Tk*LacZ* was constructed by subcloning the 5.2Kb fragment from BSKS2+; the 5.2Kb fragment was released using *HindIII* sites and end filling with T4 polymerase.

The pTk*LacZ* vector was prepared using the XbaI site, which is located in front of the Tk promoter and end filling with T4 polymerase and blunt end ligation. Orientation of the insert was checked by sequencing.

The 2.1Kb mCdx1Tk*LacZ* was constructed by subcloning a 2.1Kb fragment from the 5.7Kb BSKS2+; this 2.1Kb fragment (-3771 to –1671bp) was released using *HindIII/PvuII* sites and end filling with T4 polymerase. The pTk*LacZ* vector was prepared using the XbaI site, end filling with T4 polymerase and blunt end ligation. Orientation of the insert was checked by sequencing. The 1.4Kb mCdx1Tk*LacZ* was constructed following the same strategy, a 1.4Kb fragment from the 5.7Kb BSKS2+; was subcloned releasing the fragment using *BamHI* (-1112 to –2582bp), end filling with T4 polymerase. The pTk*LacZ* vector was prepared using the *XbaI* site, blunt end ligation and sequencing to check orientation.

The 1Kb mCdx1Tk*LacZ* was constructed by producing a 1Kb PCR fragment (+79 to –921bp), primers: RPP1Cdx1 5'-ata ctc gag cat gct gtt gcc tgg cgc cg-3' and RP1Cdx1 5'-ata ctc gag att gcc cac gta cat ggt ga-3' (Fail safe PCR kit buffer I). The PCR were run in a 1X TAE agarose gel and the band excised and electroeluted. The DNA was then phenol /chloroform extracted and precipitated with ethanol, end filling with T4 polymerase. The pTk*LacZ* vector was prepared by removing the Tk promoter using *HindIII*, end filling with T4 polymerase, blunt end ligation and sequencing to check orientation and sequence (Appendix section 2B).

4.3 Results

4.3.1 Sequencing of the Mouse *Cdx1* upstream region.

The mouse *Cdx1* upstream sequence was subcloned and sequenced. The joined sequence comprises 8.9Kb upstream of the gene, which is the actual distance to the neighbouring 5' gene. The total upstream sequence was compared with the sequence provided from the mouse Ensembl database; it showed 98.99% identity. The obtained sequence was used for consequent comparative genomic analysis, e.g. to look for putative transcription factors binding sites (TFBS) and create reporter constructs to screen for cis-regulatory elements in the *mCdx1* gene.

4.3.2 Conserved regions in the 5' region of the *Fugu Cdx1*

The *Fugu rubripes* genome has been a useful model for comparative genomics, not only in the search for new gene structures but also in the identification of conserved non-coding sequences with enhancer or regulatory value. As described in Chapter 3, the *Fugu Cdx1* non-coding sequences were extracted from the *Fugu* genome database mayffold M001324. The upstream sequences of the mouse and human *Cdx1* were extracted from the Ensembl database. For the three species, the length of the upstream sequence used for the comparisons was 12Kb.

In the *Fugu Cdx1*, the upstream region between the *frCdx1* and the neighbouring *frPdgfrβ* is 11.946Kb (see Chapter 3). Because the 5' UTR of the *frPdgfrβ* has not been mapped yet, we used a comparative analysis approach to look for the *frPdgfrβ* 5'UTR. The mouse and human *Pdgfrβ* 5'UTR were used as base sequences for the comparison. The position of the human 5'UTR was used to anchor the sequences. The Vista and Mlagan programs showed 57% level of conservation between the mouse and human *Pdgfrβ* 5'UTRs. However, the programs failed to predict a conserved 5'UTR in the *frpdgfrβ* gene. The alignment using the *Fugu* sequence also shows the absence of conserved non-coding regions upstream of the *frCdx1* gene (Figure 4.2).

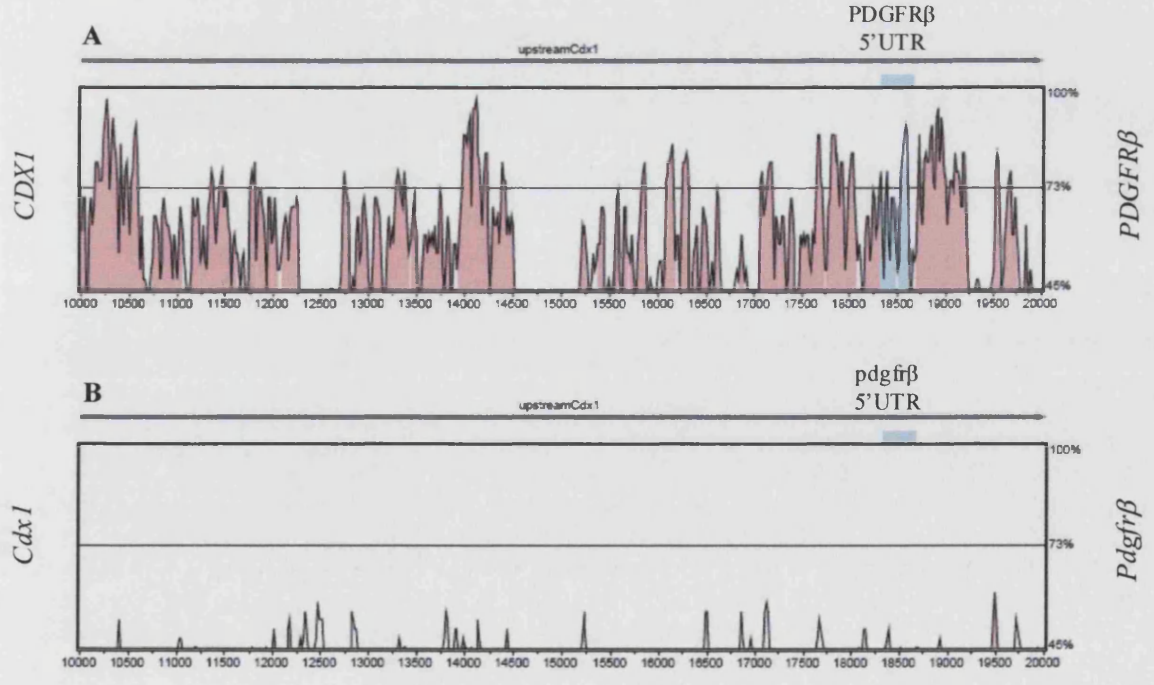


Figure. 4.2. Vista plot of the *Cdx1-Pdgfrβ* upstream region. (**A**) A plot using the human upstream region as the base sequence for comparison with the mouse. The plot defines the conserved region corresponding to the *Pdgfrβ* 5'UTR (blue colour) and the conserved non-coding regions (pink colour). The start of the *Cdx1* gene is at the left end of the plot, the *Pdgfrβ* starts at the right end. (**B**) Vista plot using the *FrCdx1* region as base sequence for comparison, no conserved 5' UTR and conserved non-coding sequences were identified by the Vista and Mlagan programs.

A second alignment of the upstream sequences of the *Cdx1* gene was done using the mouse and *Fugu* sequences to create one pairwise alignment, and the human and *Fugu* sequences to create a separate pairwise alignment. The Vista and Mlagan programs predicted small conserved non-coding sequences; each set of sequences were different in each plot. A default pairwise alignment (due to the high level of conservation between the two species) was done using the mouse and human sequences (Figure 4.3). We concluded that if the conserved non-coding regions predicted between mouse and *Fugu* and between human and *Fugu* are present in the pairwise alignment created by mouse and human, the former predicted sequences must be conserved among the three species. Each predicted element was linked to its homologue element to create a map of the conserved non-coding elements present in the *Cdx1* upstream region in mouse, human and *Fugu* (Figure 4.3).

A total of 15-conserved non-coding sequences were identified among the three species. All the predicted conserved regions are distributed evenly in the upstream regions of the genes without any special pattern or formation; the length of each element varies from 9 to 63bp (Table 4.1, Figure 4.3). However, the predicted conserved regions seem to be too small in length to be considered as conserved regulatory regions with putative regulatory characteristics. Most of the conserved non-coding regions with regulatory meaning have been identified between 500bp and several hundred base pairs long (Bejerano *et al.* 2004; Woolfe *et al.* 2005). The homology in the sequences predicted could be created just by chance without any significant meaning.

Nevertheless, the two main features to be noticed in this alignment experiment are: firstly, in the mouse conserved elements, the most distal element predicted was at

8.6Kb, relative to the translation start site. This means that no conserved regions were predicted further than 9Kb, which is the actual length of the *mCdx1* upstream region. Secondly, the first Kb upstream of each gene showed a well-conserved set of non-coding elements, located almost in the same positions (Figure 4.3).

Conserved element	Comparison	Position of the conserved element			length (bp)
		Fugu	mouse	human	
1	f/h	1953-1990	3396-3847	1793-1830	38
2	f/h	2432-2468	4899-5021	2338-2375	24
3	f/m	2287-2328	3293-3333	1249-1351	37
4	f/m	3660-3688	4774-4800	2237-2288	9
5	f/m	3820-3872	4929-4982	2351-2478	54
6	f/h	4819-4906	6201-6260	4218-4296	41
7	f/m	4917-4955	6122-6161	4125-4158	10
8	f/h	5193-5231	6373-6478	4509-4548	9
9	f/m	5998-6063	7353-7416	5984-6030	29
10	f/m	8424-8461	9365-9401	8297-8407	38
11	f/m	9459-9498	10135-10174	9529-9685	25
12	f/h	9570-9626	9726-9767	9004-9061	14
13	f/m	11319-11355	11493-11526	11419-11474	32
14	f/m	11645-11683	11729-11767	11683-11719	23
15	f/h	11883-11941	11787-11823	11729-11779	63

Table 4.1. Conserved non-coding sequences identified by pairwise alignments. The table shows the 15 conserved sequences found in the different alignments, *Fugu*/human (f/h) and *Fugu*/mouse (f/m), the position of each element in their respective sequence (in base pairs) and the length of each conserved non-coding element.

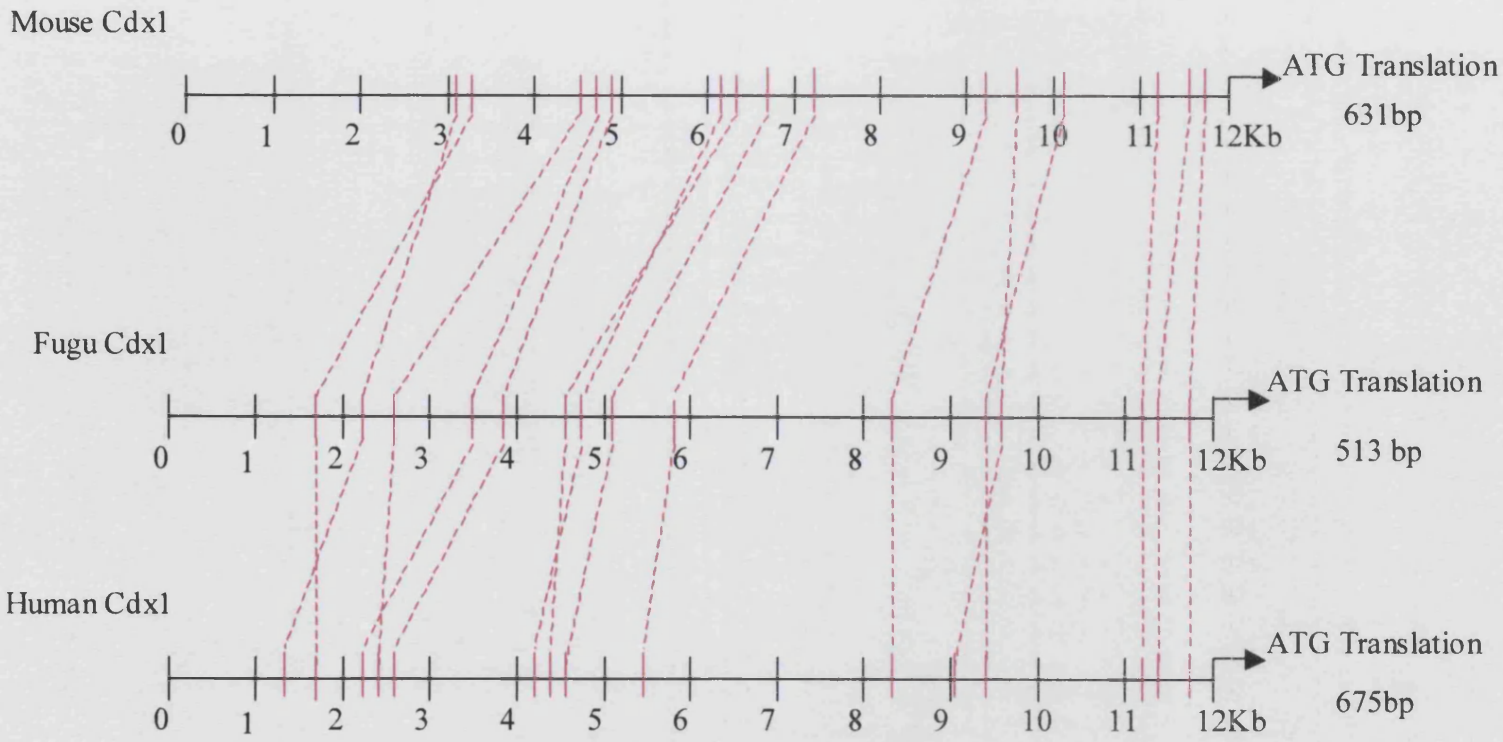


Figure. 4.3. 5' conserved non- coding regions in the mouse, human and *Fugu Cdx1* gene. Representation of the *Cdx1* upstream region in human, mouse and *Fugu*. 12Kb of genomic non-coding sequence of each gene was used in the alignment, the red bars represent each conserved non-coding region which is linked to its homologue region. The position of the first conserved element relative to the translation start site is indicated in each species.

4.3.3 Promoter comparison for the *Cdx1* gene

On the basis of the results obtained with the global alignments, and the high level of conservation present in the first Kb upstream of each *Cdx1* gene, 1Kb upstream of each sequence (*Fugu*, mouse and human *Cdx1* gene) was extracted from their respective databases; transcription factor binding site elements characteristic for promoters were predicted using the Possum program, (http://zlab.bu.edu/ ~mfrith/possum/), each sequence was screen for TATA boxes, Sp1 sites (sites rich in GC sequences) and CCAAT boxes.

The prediction of the program shows that the 1Kb upstream sequence of the *Fugu Cdx1* contains one TATA box located at -131bp relative to the translation start site (TSS) and two Sp1 sites at -443 and -816pb relative to the TSS. Using the Mat inspector program (Quandt *et al.* 1995) three *Tcf/LEF* binding sites were predicted for the sequence, one at –774bp, the second at –356bp and the third one at –39bp. The third *Tcf/LEF* site lies in the 5'UTR of the gene. No predictions for *RARE* or *RXR* were predicted for the *frCdx1* (Table 4.2, Figure 4.4).

For the *mouse Cdx1* 1Kb upstream sequence, three TATA boxes were predicted at -121, -167, and -236bp relative to the TSS position. Three Sp1 sites were predicted at positions -344, -420 and -456bp and one CCAAT box in position -674bp relative to the TSS (Table 4.2, Figure 4.5).

According to (Lickert and Kemler 2002), the 0.7Kb upstream of the mouse *Cdx1* gene is enough to drive the expression in the early mouse embryo. The authors postulate that the 0.7Kb fragment harbours two *Tcf/LEF* and an atypical RARE binding site, which are involved in the regulation of the gene

We used the MatInspector program (Quandt *et al.* 1995) to look for any *Tcf/LEF* or RARE binding sites in the 1Kb mouse upstream sequence. One RXR site (which is not a RARE site) was predicted at –83bp, which is just before the start of the transcription start site. Two *Tcf/LEF* binding sites were predicted, the first at –167bp, this site overlaps with a TATA box predicted form the Possum program; the second *Tcf*/LEF site is located at –202pb from the TSS (Table 4.2, Figure 4.5).

The 1Kb upstream sequence of the human *Cdx1* showed the presence of three TATA boxes at positions -125, -169 and –235bp relative to the TSS and two Sp1 sites at positions –156 and –567bp. Using the Mat inspector program (Quandt *et al.* 1995)

two *Tcf/LEF* binding sites were predicted at –199bp and –166bp from the TSS, the second *Tcf/LEF* site overlaps with the second TATA box predicted by the Possum program. One RXR binding site was predicted at –81bp, which lies just at the beginning of the 5'UTR (Table 4.2, Figure 4.6).

Species	TATA box	CCAAT box	Sp1 sites	RXR sites	Tcf/LEF sites
Fugu	-131bp	NP	-443, -816bp	NP	-774, -356, -39bp
Mouse	-121, -167, -236bp	-674bp	-344, -420, -456bp	-83bp	-167, -202bp
Human	-125, -169, -235bp	NP	-156, -567bp	-81bp	-199, -166bp

Table 4.2. Analysis of the 1Kb upstream sequence of the mouse, human and *Fugu* Cdx1 gene, the sequence was analysed for the presence of conserved TFBS. The table summarizes the position of each binding site in the sequences. NP: No predicted site.

Using the THEATRE program (Edwards *et al.* 2003) one TATA box was presented in the three sequences after alignment, in *Fugu* at –136bp, in mouse at –126bp and in human –123bp. These predictions match with the predictions from the Possum program, suggesting that the first TATA boxes present in the sequence may be essential for the minimal promoter of the *Cdx1* gene across the species

The analysis of the 1Kb upstream of the *Cdx1* gene showed conserved transcription factor binding sites between human, mouse and *Fugu*. The *Fugu* 1Kb region showed three *Tcf/LEF* binding sites, two Sp1 binding sites and a TATA box; no CCAAT is present in the *frCdx1*. The human and mouse 1Kb *Cdx1* region showed almost the same conservation of binding sites and positions except that the human does not contain a CCAAT box. The two human Sp1 sites are located almost in the same position to the *Fugu* Sp1 sites. The Figures 4, 5, and 6 show the actual sequences with the predicted transcription factors binding sites and the position of each one relative to the translation start site.

>frprom
GTATCGGTGCACTTTTGACCAAACATGCATTTTTAATGCATTTGGTACTCA
GATGCAGCAAAACTCATTGACACAAACTCAGCAACAGAGAGTATGTGAAT
AAACCGTTTATAATTTATGCGGACAAGTTTGCATGACAAGTTTCCGATCTG
GTTCAGCTGTCAATCGTTGGCATTATTCTGCCCCCCTGTGGTCAATCTGGG
AAATGACACGGACGAAACAAAGGCAACTTTACTTGAACATTTATCAGTCA
TATATTTACCAATGACACATTATAGAAATTAGATATTCACAGAATAAGAA
AATGTTTCTTTATTTAGCACTCAGTGTTGTCTCTCTACATCGTCACAAAGCT
GTAAAAGCATTCAAAATAATACTAAAATTGATAATAAATGATGGTTAACA
AAAACTCTAATAATA
CTCGGATGCATTTTGCAGTTGTTTGAATAAATCAAATTTGGCCATTTCAAT
GCAGCGTCGTTAAACAGAAGAGAGGAGAAGTGAAATCTTTTCAATATTAA
TGTGGTAAAAGTAAATGCTTCCGATATCCGAGGTGGGATTTATGACTTGC
ATGTTGCTATCAATAATATTAATCTTTCATGTTTAGTGTATCTGTCTACATT
GAAGTGAAGGTTAAACAAAGCTCCCCAAAAGCATTTATTTCGAGAAATGT
TTAGGGAAGAATAGGTTTTATTGTCCGTCACATGAGTCTAGCATGATAGG
CTTTATCAAAATATTCCAGCGGGCAGTAAAGCGAGAATCCCGGTCAGTAC
CTTAGCGTGGCTCCAAGACAAGCAGGTGGGTGTCAGATACCTGGGTGACG
TCACGGTGGTACTTTTG
AACTCCTCGGCAGTAAATCTATAAAGGTGTGTAAAGTCTCCCTGCACAGG
GAATCTGAACCTTAGACTGGAGTGAGCAGCGGGAATGCTGGGAGACC
TGACATGATCCCAGCTCTTTGATTTTTGTCCAGTGTCACATTTT
ACGCACCAGGACGCAGGAAAA**ATG**

Figure. 4.4. 1Kb upstream of the *Fugu Cdx1*. The underlined sequence shows the 5'UTR of the gene; sequence in red colour show the TATA box and the sequence in green colour represent the Sp1 sites predicted for the Possum program. The sequence in brown shows the *Tcf/LEF* binding sites, the first *Tcf/LEF* site overlaps with the 5'UTR of the sequence. These binding sites were predicted using the MatInspector program

>mmprom
ACTCGGGCTGATGTGTGTAGGGGTAAAACGACATGAAGCCTGGAATTTGT
TTTTAAATTGTTCCGAAAATGCTTTAGGAAAAAAAACAAAAGGCAGTAGA
TTTAAAAAAAAAAAAAAAAAAAAAGAGGTGACAGAAACCCGATAAGTA
ATGGAAGCTGCAAGGTAGGTACACAATGCAACTCGGTGTATATGTATGTT
TGAAACATTTCAGCAATGATAATGTCaAGGTAATACGCATCaAAAGTTTGT
GAAGTTTCCCAAGGTGTCCCCAGAAACCCCCCTGGTCCACACCTTTGAGA
ATTTCTCTGGGGAGCCAATAAGAGAAACAGTAGCTTGCAGACTGGCTAT
CTGCATATGCACATCACCATAGCTTTCATCTGGGACCCGAAgGGTCGTGAC
CCCTAACCcCACCACCACCCCTCCCAAgAGACGCCCGCTACCTTACAGAC
GGaAcCCCCGTTTGAAGTCAGCCTTGCTCTCCGCCTCACCTCAAgCTGGT
GGGTCACTGCTGACAGTTGTCCCcATGCtCtGTTGGGCGGTGTgAGGCTC
gCCtAgGGtCATGCCACcACTCCACCCCGCTCCTCGGAGCCAGTTTCtCC
ACCTGTAACCCAGGGGTGGGTGGTGGGGAGGTcCtgCGACCCCCGAAGGA
GGAgTCGGAACCCCAAGCTGGGACCGGACCGTCATCCAGGCCGCGCCCTC
CGGGTCCCCAGCCCGCTGGCCCATCCACCTCCCGCTTAGGGCGGCAATTT
GTCTCCTTTTGAACCCCCTCGCCCGACGGGCTTCCCCCTTTGATTCGCGG
CCCCGAGGCTTCCCCCCGCTTTGAAATGCAAAGCCGCCCGGCTGGGGCCG
CGGACGGCCCGCGGCTATAAAAGGCCGGGGTGGGGCGGGCGCGGCGGCG
GCCGCGGGGCACAGGTGAGCAGTCGCTGGTCGTCGGGGCGGCTCGCTCGG
GCGCGGCGGCGCCAGGGCCCAGCATGCGCGGGGGACCCTGCGGTCACCA
TG

Figure. 4.5. 1Kb upstream of the Mouse *Cdx1*. The underlined sequence shows the 5'UTR of the gene; sequences in red colour show the TATA boxes, the sequences in green colour represent the Sp1 sites and the sequence in blue colour the CCAAT box predicted for the Possum program. The sequence in brown shows one of the *Tcf/LEF* sites, the second *Tcf/LEF* site overlaps the first TATA box in the sequence. The sequence in dark green shows the RXR binding site predicted by the MatInspector program.

>hsprom
ACTTCAGAAAAGCGGGGAGTAGGTGAAATAAGTGTGACAAAACCTTAAT
AGTTGTTGAAGCGGGGTGATGGGTACACAGCAGTTCATTATACAATTCTA
CTTTTGTGTATGCTTGAAACCTCTCAATTAAAAAAAGAGGGTAGAGAGTC
ATCAAGATAATTTACTAAAAGTTTCCTGTAATGTCCCCAGAAGACCCCCCT
AACTCACACACACACACACCCACACACACACTGCCTTTTGAGAACTTTCTC
TAGAAGGCCAACCCAGGCAGTCCCCCAACTGTAAGGAAGACTCGTGTATG
TATGTGCATATGTGCATTTCCCCAGGGAAAAACATCCACAGCTTCCATGA
CGAGAGGGGTCGTGACCCCTCCCCGCCAAAAGATTAAGGACCTGCGATCC
TACAGACCGGAGCCCTGTTTGAAGTCTGCGTTGCCCCTCACCTCAAGCTGG
TCACTGTGTGAAGTTGGCCTAGAATCCCCCGGCCCCGCTGAAGTATGTAG
GGGTAAAATGACATGATGTCTGGAATTTGTTTTAAAATTGGGAGCTTGTTC
CTCCGCCTGTAAAATGGGGCTGCAGGGCCGTCCACGCGGCCACCGGAAGG
ACAAGGTGTTCAGGCCGCTAGGCCGCTCCCTGGCAAGCGATTCCCACTCG
CAGCGCGGCCTCGACCCTCGCCCAAGACGCGCCCTCCGCGCCCCCACCCC
CTCCAGGCCCTGGCCAGTCCACCTCCCGCTTGGGGCGGCAATTTGTCTCCT
TTTGAACCCCCCGCCCCCGACGGGTTTCCCCCTTTGATTCGCGGCCCGGAG
GCTTCCCCCCGCTTTGAAATGCAAACCCGCCTCGGCTGGGGCCGCGGGCG
GCCCGGAGCTATAAAAGGCCTGGGTGGGGCGGGCGCGGCGGCAGGACAG
CCGAGTTCAGGTGAGCGGTTGCTCGTCGTCGGGGCGGCCGGCAG
CGGCGGCTCCAGGGCCCAGCATGCGCGGGGGACCCCGCGGCC
ACCATG

Figure. 4.6. 1Kb upstream of the Human *Cdx1*. The underlined sequence shows the 5'UTR of the gene; sequences in red colour show the TATA boxes and the sequences in green colour represent the Sp1 sites predicted for the Possum program. The sequence in brown shows one of the *Tcf/LEF* sites, the second Tcf/*LEF* site overlaps the second TATA box in the sequence. The sequence in dark green shows the RXR binding site predicted by the MatInspector program.

4.3.4 Conserved transcription factors in the upstream region of the *Cdx1* gene

Based in the global pairwise alignments and the analysis of the proximal 1Kb region, conserved transcription factors binding sites were predicted using the MatInspector and TESS programs for each of the upstream sequences. Due to the high number of transcription factor binding sites predicted for each sequence, we eliminated those transcription factors that we considered not to be important in the regulation of the *Cdx1* gene. Based on the literature, those transcription factors, which are expressed in the intestine, epithelial cells or during early development, were kept.

Each individual transcription factor or cluster of transcription factors was located in the upstream element cloned in to the GFP or *LacZ* expression vector. For

the mouse *Cdx1* constructs, the 5.2Kb fragment contains binding sites for Pax, RXR and RAR factors which are expressed in intestinal tissue as well as being involved in vertebral patterning. *Tcf/Lef* complexes and *Cdx* binding sites are also present in this element; A STAT complex binding site was predicted, the STAT factor has been involved in chromatin remodelling (Table 4.3).

The 2.1 and 1.4Kb elements harbour RAR and Est binding sites, these factors are involved in vertebral patterning and brain development. The 1Kb upstream element, besides the transcription factors already described above, was predicted with HNF, GATA and Oct binding sites, which are widespread factors (Table 4.3).

The final consensus binding sites predicted for the *Fugu* elements show a very similar repertoire of TFBS compared with the mouse elements. The 1 and 2Kb elements contain GATA, HNF, Cdx and Oct binding sites, all these factors are know to be expressed in intestinal tissue. The 4Kb element has RXR, RORA and Pax binding sites, which are very similar to the binding sites present in the 5.2Kb mouse element (Table 4.3).

The final consensus binding sites present in each construct are summarized in table 4.3. The figures 4.7 and 4.8 show the *mCdx1 LacZ* and *frCdx1 LacZ* constructs and the position of the fragment in the 5'upstream region of the mouse and *Fugu Cdx1*. The conserved non-coding regions are also shown; however, as mentioned earlier these regions were not considered for the TFBS analysis.

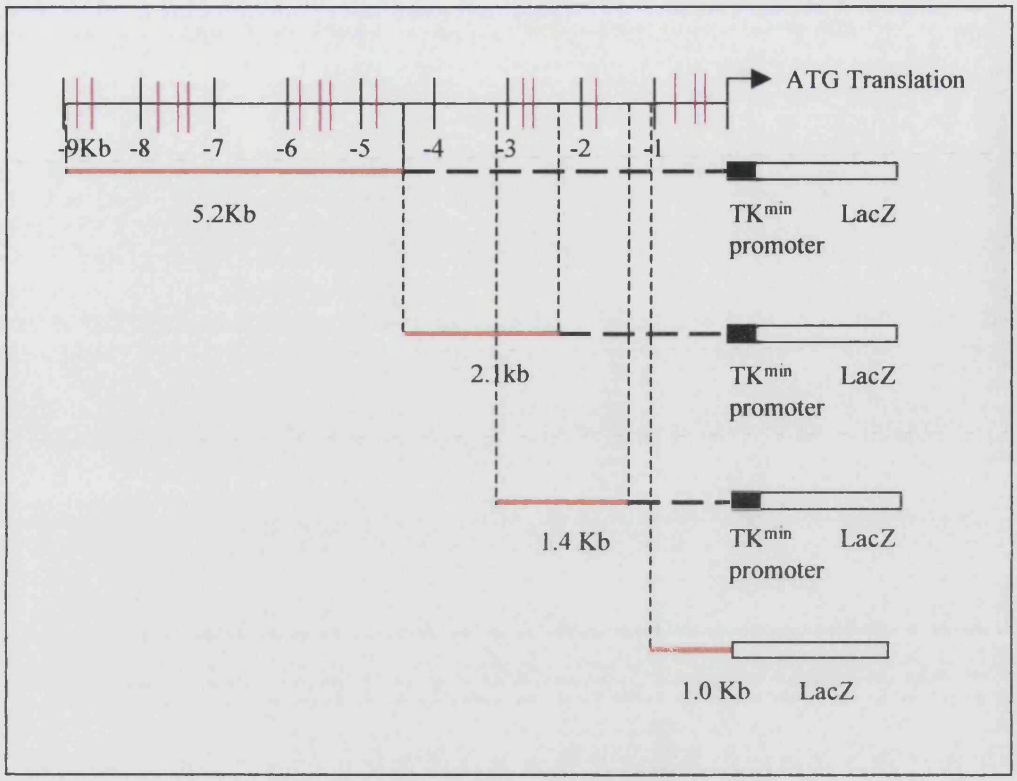


Figure. 4.7. Cis- regulatory elements in *mCdx1LacZ* reporter constructs. Orange lines show the upstream fragments fused to the reporter vector, dotted lines show where the cis- elements lie in the upstream sequence and the distance between the translation start site and the element.

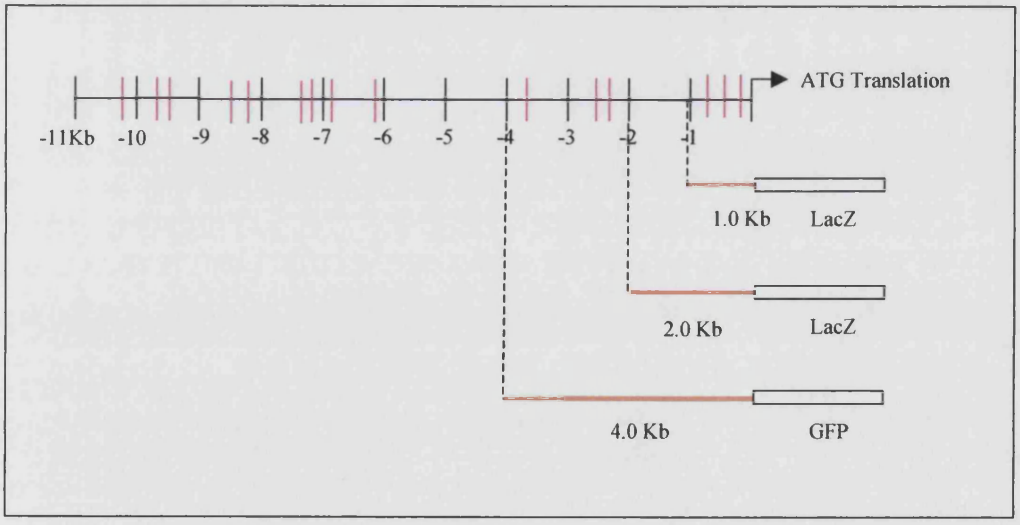


Figure. 4.8. Cis- regulatory elements of *frCdx1 LacZ* reporter constructs. Orange lines show the fragments fused to the reporter vector, dotted lines show where the cis- elements lie in the upstream sequence and the distance between the translation start site and the element.

construct	TF complex	TF name	TF expression
5.2Kb *TkLacZ mCdx1*	cEST/GATA		
	GATA/TCF-LEF/Pax1		Pax1 expressed in developing vertebral column
	FoxA/OCT1/Thing1	FoxA, Fork head domain factor similar to HNF	Thing1 expressed in a variety of embryonic tissues
	STAT/STAT	signal transducer and activator of transcription from the C/EBP family	activator factor, involved in chromatin remodelling
	GATA/PDX1	Pdx1, pancreatic homeodomain factor	expressed in the intestine
	cEST/RORAa/HNF4/GATA/FOXL1	RORA, nur subfamily of nuclear receptors	expressed in the intestine
		FOXL1, similar to HNF	
	OCT1/AP/Thing1	AP, activator portein	
	RORA/T3Ra	T3R, thyroid hormone receptor	
	RAR/CAAT	RAR, retinoic acid receptor	vertebral patterning, differentiation and growth of epithelial cells
	TATA/CDX2	TATA, basic binding site for minimal promotors	CDX family expressed in the intestine
	TCF/LEF		
	FOXD1/RXR	FOXD, similar to HNF4	
		RXR, retinoic X receptor	differentiation, growth and homeostasis of epithelial cells in various tissues
	RXR/RXR/AP	AP, activator portein	
	TCF-LEF/GATA	TCF, trans-acting T-cell factor	
	TCF-LEF/TATA/CDX2		
2.1Kb *TkLacZ mCdx1* & *1.4Kb TkLacZ mCdx1*	RARα	RAR, retinoic acid receptor	vertebral patterning
	TCF-LEF/CCAAT: CCAAT	CCAAT, enhancer binding protein beta	
	ER	estrogen receptor	
	CEST/GATA	EST, lymphoid differentiation factor	Est1 & Est2 expressed in developing brain, gut and skin, involved in prolifferation and differentiation
	ERE/TCF11/AP1	ERE, estrogen response element	
1.0Kb *LacZ mCdx1*	HNF3/HNF4/GATA	HNF, hepatic nuclear factor	expressed in epithelial cells and endoderm and mesoderm derivied organs
	GATA		GATA, widespread expression
	FOXF2/OCT1/OCT1		
1.0Kb *LacZ frCdx1* & 2.0Kb *LacZ frCdx1*	TCF/LEF	LEF, lymphocyte factor	forms a complex with β-catenin, expressed in intestine
	HNF4		expressed in the intestine
	LEF1/RXR/HNF4		
	TATA/CDX2	TATA box, Caudal factor	
	Pax3	paired domain protein	expressed during embryogenesis
	OCT1/FOXC1		
	TCF11/TCF-LEF		involved in the Wnt signal transduction pathway
	GATA/HNF1/FOXA2/ CCAAT		GATA/HNF complex expressed in the intestine
	GATA/PDX1		
	CDX	Caudal factor	expressed in early development and intestine
	CREB	cAMP-response element binding protein	
	GATA/OCT1		
	OCT1/OCT1/OCT1	TF from the POU domain family	neural fate specification
4.0Kb *LacZ frCdx1*	RXR		
	RXR/Ap1	Ap1, activator protein	
	Pax1		expressed in the developing vertebral column
	Pax3		expressed during embryogenesis
	Pdx1	pancreatic and intestinal homeodomain factor	
	Cdx/RORA	RORA,nur subfamily of nuclear receptors	RORA expressed in the intestine
	Sp1	stimulating protein	widespread expression
	CDX/TATA/HNF1/Oct1		

Table 4.3. Representative transcription factors binding sites contained in each of the expression constructs for the mouse and *Fugu Cdx1*. Individual or complexes of transcription factors are shown with the place where they have found to be expressed (Gaub *et al.* 1990; Tronche and Yaniv 1992; Kane *et al.* 1995; Lopez-Rodriguez *et al.* 1997; Singh *et al.* 1997; Dussault *et al.* 1998; Emami *et al.* 1998; Larsson *et al.* 1998; Lee *et al.* 1999; Allan *et al.* 2001; Smits *et al.*

2001; Consales and Arnone 2002; Flock and Drucker 2002; Jacobsen *et al.* 2002; van Heel *et al.* 2002; Almeida *et al.* 2003; Afouda *et al.* 2005; Moore-Scott and Manley 2005).

4.3.5 Reporter assays in CaCo2 cells

To investigate which are the regulatory regions involved in the regulation of the *Cdx1* gene, reporter assays were performed in Caco2 cells. Transfection of the reporter constucts shows that the *mCdx1 LacZ* 5.2Kb element, which is situated in the distant end of the 5' upstream region, presents a 42.6 fold induction when compared to the *pLacZ* vector. The 2.1Kb and the immediate 1Kb element were also able to drive the expression of the reporter with 31.5 and 3.6 fold induction respectively. The 1.4Kb element did not show induction of the reporter. The *Fugu Cdx1 LacZ* constructs were also able to drive the expression of the reporter in CaCo2 cells. The 1Kb upstream of the gene showed 37.3 fold induction, and the 2Kb element 3.8 fold induction when compare to the basal activity of the *pLacZ* vector (Figure 4.9).

The results of the transfection assays indicate that the mouse and *Fugu* upstream regions contain regulatory elements able to drive the expression of the reporter in intestinal cells. The 5.2Kb element of the mouse *Cdx1*, which is located in the distal region of the gene, drives the expression of the reporter as high as the positive control (pNLS*LacZ*), which is the highest expression of all the mouse constructs, indicating that this region is possibly an intestinal enhancer of the *Cdx1* gene.

The 2.1 and 1.4Kb are two elements that overlap in the upstream region of the gene (see Figure 4.7), the 2.1Kb element is able to drive the expression of the reporter with a very decent induction; this is not the case for the 1.4Kb element, which shows a null induction of the reporter, suggesting the presence of a repressor element in the overlapping region or in the complete 1.4Kb region. The 1Kb upstream of the mouse *Cdx1* is also able to induce the expression of the reporter; this element is one of the most conserved in terms of non-coding regions and transcription factors with the *Fugu* and human *Cdx1*, suggesting that the basic machinery to induce the expression of the gene is contained in this region in the three species.

The transfection result also shows expression of the *frCdx1-LacZ* constructs in CaCo2 cells. In contrast with the 1Kb mouse element, the *Fugu* 1Kb upstream region

shows very high induction of the reporter, indicating that not only the minimal promoter is contained in this element but also some elements which enhance the expression of the gene. The 2Kb *Fugu* element that extends from the TSS is able to drive the expression of the reporter but not as high as the 1Kb element, suggesting the presence of a repressor in the second Kb of this region.

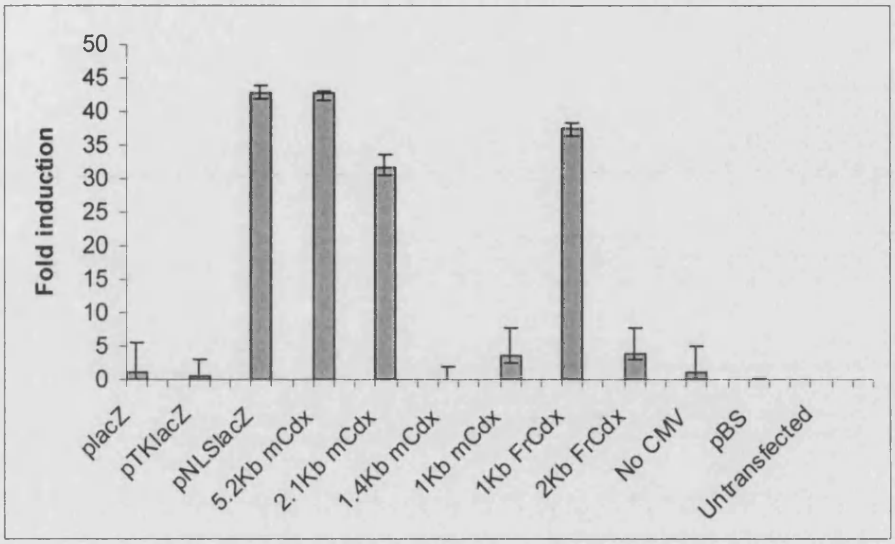


Figure. 4.9. Activation of the mouse and *Fugu Cdx1* upstream regions. CaCo2 cells were transfected with *mCdx-LacZ* and *frCdx1-LacZ* constructs, pNLsLacZ was used as a positive control and pBSKS2+ (pBS) as a negative control for transfection. Each bar represents the average fold of induction of normalized values ± range bars of triplicate points of 2 experiments.

4.3.6 Analysis of the *frCdx1* cis-regulatory elements in the developing zebrafish embryo

To test if the upstream region of the *Fugu Cdx1* gene is sufficient to control the expression of the gene in early and late stages of development, three GFP expression constructs containing different *frCdx1* non- coding sequence lengths were constructed. Expression constructs were injected into the 1-2 cell-stage embryo. Embryos were analysed at 5, 24 and 48hrs post- injection (hpi) for GFP expression.

The *Fugu Cdx1* expression constructs were injected at three different concentrations, 15, 25 and 50ng/µl. The 15 and 25ng/µl concentrations showed a similar percentage of GFP expression for the three constructs. The 50ng/µl

concentration produced a high level of mortality to the injected embryos. The few survivors showed severe malformations after the gastrulation stage (Table 4.4).

construct	concentration	GFP exp/ #embryos injected				
		80% epiboly	tissue expression	14 somites	2 dpf	4 dpf
4KbfrCdx1	25ng/μl	60/96	tail bud	12	9	4
			distributed exp	16	13	7
			SC/NC	0	0	0
			total	65	61	61
4KbfrCdx1	15ng/μl	34/60	tail bud	4	2	2
			distributed exp	5	0	0
			SC/NC	0	0	0
			total	26	24	24
2KbfrCdx1	25ng/μl	41/45	tail bud	7	0	0
			distributed exp	12	9	6
			SC/NC	0	0	0
			total	27	27	27
2KbfrCdx1	15ng/μl	28/43	tail bud	8	2	2
			distributed exp	8	11	9
			SC/NC	0	0	0
			total	31	20	18
1KbfrCdx1	25ng/μl	64/70	tail bud	10	10	6
			distributed exp	15	7	7
			SC/NC	0	0	0
			total	50	48	32
1KbfrCdx1	15ng/μl	26/48	tail bud	9	8	7
			distributed exp	12	15	14
			SC/NC	0	0	0
			total	45	40	36
2KbfrCdx1	50ng/μl	Most of the embryos were dead by 80% eoiboly. The survivors embryos showed malformations				
1KbfrCdx1	50ng/μl	Most of the embryos were dead by 80% eoiboly. The survivors embryos showed malformations				

Table 4.4. Expression of the *frCdx1*-GFP reporter constructs in to the zebrafish embryo. Embryos were screened for GFP expression at 80% epiboly, 14 somite stage, 2 and 4dpf. Expression was observed in tail bud and distributed expression along the embryo. The spinalcord (SC) and notochord (NC) did not show GFP expression.

A GFP reporter construct carrying the *frCdx1* upstream from position –4000 to +81bp relative to the transcription start site was designed to look for the regulatory elements involved in the regulation of the gene. Expression of the transgene was

observed at 80% epiboly; 64.4% of the injected embryos showed expression at this stage. At the 14 somite stage, 16 out of 91 embryos showed GFP expression in the tail bud region. By 2 and 4dpf, the number of embryos expressing the transgene in this region decreased to 11 and 6 embryos respectively. A distributed expression was detected at the 14 somite stage, 21/91 embryos showed this expression. By 2dpf 13/85 embryos retained this pattern of expression and by 4dpf, only 7/85 embryos showed expression. This distributed expression was mostly present in muscle cells, mesodermal cells and neural crest cells. No expression was detected in the spinal cord (SC) or notochord (NC) in any of the three stages screened (Table 4.4).

A second *frCdx1*: GFP construct containing -2000 to +81bp relative to the transcription start site was generated to look for regulatory regions. Injection of this construct into the zebrafish embryo produced the same pattern of GFP expression obtained with the 4Kb*frCdx1*: GFP construct; 78.4% of the embryos (69/88) of the embryos expressed the reporter at 80% epiboly; 25.86% of the GFP expressing cells were located in the tail bud region at 14 somite stage. By 2 and 4dpf, only two embryos remained with this pattern of expression. A similar GFP distribution pattern was observed with this construct, 34.5% of the embryos (20/58) showed expression at 14 somite stage. By 2dpf, 42.5% of the embryos (20/47) remained with this expression and by 4dpf, 33.3% of the embryos (15/45) maintained the same expression pattern. No GFP expression on the notochord (NC) and spinal cord (SC) was detected at any stage with the 2Kb*frCdx1*: GFP construct (Table 4.4).

A smaller *frCdx1*: GFP construct expanding from –1000 to +81bp relative to the transcription start site was constructed to investigate whether the cis-regulatory elements and the promoter are contained in the first Kb proximal to the gene. Transient zebrafish embryos expressing the 1Kb*frCdx1*: GFP construct resembled the same pattern of expression obtained with the two anterior constructs. Expression of the reporter was first observed at 80% epiboly stage, 76.3% of the embryos (90/118) showed GFP expressing cells. By the 14 somite stage, 20% of the embryos (19/95) showed GFP expression in the tail bud region. At the same stage, 28.4% (27/95) showed a distributed GFP expression. By 2dpf, 20.4% of the embryos (18/88) retained the expression in the tail bud region and 25% (22/88) retained the same distributed expression. Expression of the transgene was still present at 4dpf, 19.1% of the embryos (13/68) kept the tail bud region expression and 30.8% (21/68) of the

embryos showed distributed GFP expression. No GFP expression on the notochord (NC) and spinal cord (SC) was detected at any stage with the 1Kb*frCdx1*: GFP construct (Table 4.4). The distributed GFP expression was present mainly in muscle cells, mesodermal cells and neural crest cells, a minority of blood cells and heart muscle cells showed expression, few cells near the gut expressed GFP. The transient GFP expression analysis of the frCdx1 reporter constructs into the zebrafish embryois shown in the appendix section 5A.

4.4 Discussion

I have analysed and characterized the upstream non-coding region of the mouse and *Fugu Cdx1* in search of conserved regulatory elements involved in the regulation of the gene. The reporter assay data indicate that specific regions of the *mCdx1* and *frCdx1* are able to produce reporter expression in intestinal cells. Furthermore, expression of the *frCdx1* elements in transgenic zebrafish are able to reproduce, to some extent, the expression of the wild type gene.

Using a comparative genomics approach, we analysed and compared the upstream region of the human, mouse and *Fugu Cdx1*. As discussed in Chapter 3, there is a high level of conservation in the coding region of the genes; however, no conserved non-coding sequences were identified in the upstream region of the gene. Further analyses using the intronic regions and the 3' downstream region showed the same result. Small non-coding sequences were conserved among the three species; however, the functionality and significance of these non-coding regions remains elusive.

Unexpectedly, a search of transcription factor binding sites in the upstream region of the gene showed a conserved repertoire of binding sites. The region that shows a high level of conservation is the first Kb upstream of the gene. The three species contain a TATA box in their promoter, and Sp1 sites, which are characteristic elements for the basic transcription machinery of gene. Remarkably, the three species contain a *Tcf/Lef* binding site in this region. Previous reports suggested that the mouse *Cdx1* is controlled by a *Tcf/Lef* element located in the first -700bp of the gene (Lickert *et al.* 2000), which suggests that the trans-acting factor involved in the *Cdx1* expression has been conserved across species and during evolution.

Further analyses of the upstream region showed a defined group of conserved transcription factors along the sequence. RARE and RXR sites were predicted in both sequences, either in distal or proximal position. Previous studies indicate that RARE is necessary for the expression of *mCdx1* and, RARE and RXR factors are crucial for normal vertebral patterning (Houle *et al.* 2003). Biochemical and embryo culture studies have found that *Cdx1* autoregulates its expression, although this regulation may be mediated via LEF1 factor and not by direct protein/DNA interaction (Prinos *et al.* 2001; Beland *et al.* 2004). The analysis of the upstream sequences also shows the presence of Cdx and Tcf/Lef in the distal region of the gene, suggesting a conserved mechanism in the regulation of *Cdx1* across species.

The data obtained from the transfection assays suggests that the TFBS predicted for each of the elements might be playing an important role in the activation of *Cdx1*. The mouse 5.2Kb fragment possibly contains enhancer regulatory sequences important for the intestinal expression; important factors such as CDX, GATA, LEF and STAT, which are expressed in the intestine, are contained in this fragment. The fold induction obtained for the overlapping 2.1 and 1.4Kb elements suggest that some of the transcription factors in these two regions, especially in the 1.4Kb fragment, may act as repressors or co-repressors of the *Cdx1* expression. This may be also the case for the *Fugu* 2Kb element in which a reduced fold induction was obtained. Although the TFBS predicted have been described as activators, some may act as repressors when combined with specific factors. The 1Kb upstream element showed activity in intestinal cells in both species. As the anterior reports and the TFBS analyses suggest, this region contains the necessary elements to activate the gene in intestinal cells in both species.

Comparative genomics studies have found clusters of conserved non-coding sequences across the genome, mainly located in introns or in flanking regions of regulatory genes involved in development. These conserved sequences with suggested regulatory characteristics were not found near the *CDX1* gene (Bejerano *et al.* 2004; Woolfe *et al.* 2005). However these studies do not answer how some regulatory genes, which also participate in development and have no significant level of conservation in their flanking sequences, are regulated.

The absence of conserved non-coding regions nearby the coding sequences is not an uncommon characteristic of the genes; a clear example is the regulatory region

of the *even–skipped* gene in *Drosophila melanogaster* and *Drosophila pseudoobscura*; these two species shared a common ancestor from 40 to 60 million years ago, and their cis-regulatory region drives the same detail of expression pattern in both species despite the underlying sequences being highly dissimilar. Even though the regulatory sequence of the ancestral *even-skipped* gene has gradually changed in the two fly species, the functional activity of the cis-element has not changed (Boffelli *et al.* 2004).

The deletion promoter analysis using the upstream region of the *frCdx1* showed that the upstream region of the gene controls early expression in the zebrafish embryo. The earliest expression identified in the transient embryos was at 80% epiboly, which is slightly later than the *in situ* studies reported for the *ZfCdx*, where expression starts at 50% epiboly, but is evident at 80% epiboly (Joly *et al.* 1992). The three expression constructs extending from +1 to 4Kb were able to drive the expression of the reporter at this stage, indicating that the 1Kb upstream of the translation start site contains the elements necessary to regulate the expression of the gene at this early stage. The three constructs showed the same level of GFP expression. Previous studies have not been able to identify the *Cdx1* regulatory regions involved during the gastrulation period.

By the 14 somite stage, the *frCdx1* reporter constructs were mainly expressed in the tail region. This agrees with the *ZfCdx* and the *mCdx1* expression reported for the same stage; however, both genes are also expressed in other tissues. The *ZfCdx* is also expressed in spinal cord and notochord (Joly *et al.* 1992). In mouse, *Cdx1* also shows expression at the level of the posterior hindbrain, dorsal margin of the neural folds, anterior and posterior somites, and in the mesodermal cells of the forelimb and hindlimb buds (Meyer and Gruss 1993). Even though we were unable to identify expression in the notochord, spinal cord and developing neural tube, some of the distributed GFP expression was in mesodermal cells (which may form part of the notochord) and in the neural crest cells. This agrees with the mouse expression at 8.75dpc, when the embryo is around the 14 somite stage.

By the 2 and 4dpf, the *frCdx1*-GFP expression remained unchanged; the tail bud still showed expression of the transgene and no notochord or spinal cord expression was detected. By 22hpf, the *ZfCxd* is expressed in neuroectodermal cells, spinal cord and tail bud; after 48hpf, no expression has been detected (Joly *et al.*

1992). In the mouse, by 10dpc the embryo has developed to the 30 to 34 somite stage, which corresponds to the 22-24hpf stage in the zebrafish. At this stage the *mCdx1* expression is located in the forelimb and hindlimb bud and in the tail bud. The latest expression is seen at 12dpc in the somites (Meyer and Gruss 1993). The results obtained still agree with the wild type expression of the gene in the tail bud region. Because the GFP protein still remains active for a period of time, it is difficult to know exactly when the activity of the transgene finished.

It is possible that the regulatory elements involved in the expression of the gene in these specific tissues are located in the intronic regions, in the 3' region or even further up of 4Kb in the 5' region of the gene. Studies performed by Gaunt et al. (2003) using the chick *Cdx1* (*CdxA*) non-coding region demonstrated that the gene requires 1.4Kb of the upstream region and 2.1Kb of the first intron to drive the expression of the reporter in the neural tube and in the first somites (E8.5). Further analyses of the non-coding *CdxA* sequence revealed the presence of one RARE site and a *Tcf*/β-catenin binding site in the intronic region of the gene; independent disruption of either of the sites causes a down regulation or lack of specific expression in the neural tube and somites.

We were also unable to identify expression of the trangenes in the intestinal tissue. By 5dpf the zebrafish intestine becomes active, but is not until 7dpf that most of the genes are expressed in this tissue (Wallace *et al.* 2005). Further analyses at later stages will indicate if the *frCdx1* elements are active in the zebrafish intestine and if they show a gradient of expression.

Further investigation of the actual mechanism that governs the expression of the *frCdx1* and *mCdx1* will have to be carried out. Of special interest will be to look for the cis-regulatory elements responsible for neural tube and spinal cord expression. More detailed studies, for example, mutation of specific binding sites, or deletion of specific regions, will define which are the factors involved in the regulation of *Cdx1*. Finally, transient zebrafish carrying the *mCdx1* upstream elements will allow identification of the regions involved in the expression of the gene at different developmental stages.

The data presented here, although preliminary, indicate that the *frCdx1* 1Kb contains the elements necessary to activate the expression of the gene at early stages

of development, during development and later in the intestine. The distal region of the *mCdx1* contains enhancer elements important for the activity of the gene in intestinal cells. The TFBS present in the *Cdx1* upstream region and especially the binding sites located in the proximal upstream region, strongly suggest that the regulatory mechanisms that govern *Cdx1* have been preserved through evolution.

Chapter Five

Identification of the *Apc* genes in *Fugu rubripes*

5.1. Introduction

One of the main roles of the Adenomatous Polyposis Coli protein is to contribute to the formation of a phosphorylation complex that regulates β-catenin in the Wnt signalling pathway. APC also contributes to the migration of intestinal cells in the crypt-villus axis; over expression of APC results in an altered migration of intestinal cells (Mahmoud *et al.* 1997). Most studies of this protein address its contribution to the development of colon cancer; mutations or complete loss of APC have been found to precede the formation of polyps, which are the precursors of adenomas and adenocarcinomas in colon (Fodde 2002).

There are two types of specific domains contained in APC, the domains near to the N-terminus of the protein, and the domains in the C-terminus of the protein. The oligomerization domain and the armadillo repeats are located in the N- terminus of the protein. The 15aa repeats and the 20aa repeats are located in the central portion. The basic domain, the EB1 and human disc large (HDLG) binding domain are situated in the C-terminus of the protein.

The heptad repeats in the N-terminus permit APC to form homo- dimers (Su *et al.* 1993). The armadillo domain is constituted by seven repeats; it is highly similar to the β- catenin domain. This region also binds to the phosphatase 2A protein (PPA2), an enzyme that also binds to Axin (Hsu *et al.* 1999; Seeling *et al.* 1999). This domain also binds to the APC- stimulated guanine nucleotide exchange factor (Asef). The APC/Asef complex activates the Rac and Rho GTP binding proteins, triggering a mechanism involved in the stabilization and motility of the actin cytoskeleton scaffold (Kawasaki *et al.* 2000).

The 15aa repeats also bind to β- catenin; however, this binding does not cause its down regulation. These repeats are unique to the APC protein and there are no similar domains in the β-catenin protein (Ozawa *et al.* 1989; Rubinfeld *et al.* 1993). There are seven 20aa repeats located in the central region of APC; these repeats also bind to β- catenin after being phosphorylated by GSK3β; although the seven 20aa repeats are phosphorylated, only three are necessary for the down regulation of β-catenin (Figure 5.1). The majority of APC mutants lack the complete set or many of

the 20aa repeats, suggesting that this domain is important in tumorigenesis (Polakis 1997; Rubinfeld *et al.* 1997).

The three Axin-binding sites in APC are contained between the third and fourth 20aa repeats, the fourth and fifth 20aa repeats, and the seventh 20aa repeats and the basic domain. Axin facilitates the complex formation of APC/β-catenin and therefore phosphorylation of both proteins by GSK3β. Overexpression of Axin down regulates β- catenin. As in the case of APC, Axin acts as a negative regulator of the Wnt signalling pathway (Hart *et al.* 1998; Kishida *et al.* 1998).

In the C-terminus domain of APC lies the basic domain of the protein. This domain is rich in arginine and lysine residues (basic amino acids), and a high number of proline residues (Groden *et al.* 1991). Studies have shown that the C-terminus binds to microtubules and stimulates proliferation of tubulin *in vitro* (Munemitsu *et al.* 1994). This domain is not affected in colorectal tumours (Figure 5.1).

The EB1 and the HDLG binding domains are also contained in the C-terminus of the protein. The EB1 protein is involved in the binding to the centromere, the mitotic spindle and the distal (+) tip of the microtubules (Berrueta *et al.* 1998; Morrison *et al.* 1998). APC also associates with the + microtubules in the cytoskeleton; the removal of the EB1 domain in APC reduces the binding capacity of the protein to the + ends of the microtubules (Tirnauer and Bierer 2000). Although the majority of colorectal cancers have mutation(s) in the C-terminus, mutations in or lack of the EB1 domain in APC has not been associated with intestinal tumours (Smits *et al.* 1999). The human disc large binding domain is situated in the final end of the C-terminus in APC (Figure 5.1). The APC-HDLG complex is involved in the cell cycle progression from the G_0/G_1 to the S phase (Baeg *et al.* 1995).

Although the presence of APC in the nucleus has been confirmed (Henderson 2000), the exact position of the domain with nuclear location signal is still unclear. Three putative nuclear signals have been identified, carrying the LXXXXLXXLXL and VXXXVXXVXV motifs. Deletion of two of these motifs disrupts the binding of APC with CRM1 (a nuclear export protein), and APC accumulates in the nucleus (Henderson 2000). Rosin-Arbesfeld et. al. (2000) confirmed that APC shuttles β-catenin from the nucleus to the cytoplasm in a CRM1 depended process. This study also demonstrated the presence of three nuclear localisation signals in APC, each

signal with the repeat LXXLXL/I/M/V. These nuclear signal motifs are contained in the third, fourth and seventh 20aa repeats of the protein. Two other putative nuclear signal motifs were found, one in the N-terminus and the second after the armadillo region.

It has been also suggested that APC can bind to DNA. Three putative binding sites have been located in the protein with the S/TPXX motif; these domains bind to A/T rich nucleotide sequences, however, regulation of transcription has not been shown yet (Deka *et al.* 1999).

Inactivation of APC or direct inactivation of β-catenin by mutations in its APC binding site provoke accumulation of β-catenin in the cytoplasm. Studies have shown that introduction of wild type APC into cells with mutated APC protein reduces the pool of β-catenin (Korinek *et al.* 1997; He *et al.* 1998). Once β-catenin is inside the nucleus, it associates with Tcf/Lef to activate the transcription of genes involved in cell cycle progression, such as *c-myc*, *cyclin D1*, *conexin 43*, *metalloproteinase matrylsin* and *Cdx1*, among other genes (van der Heyden *et al.* 1997; Crawford *et al.* 1999; Shtutman *et al.* 1999).

One of the roles of APC outside of the Wnt pathway is in intercellular adhesion. The cytoplasmic domain in cadherins responsible for cell adhesion is the SLSSL, that is found in four of the seven 20aa repeats in APC (Nagafuchi and Takeichi 1988; Jou *et al.* 1995). Whereas β-catenin associates with E-cadherin in a complex to bind to the α-catenin and the actin network, APC binds to the microtubules that do not belong to the actin network and localizes at the edges of the cells (Smith *et al.* 1993; Neufeld and White 1997).

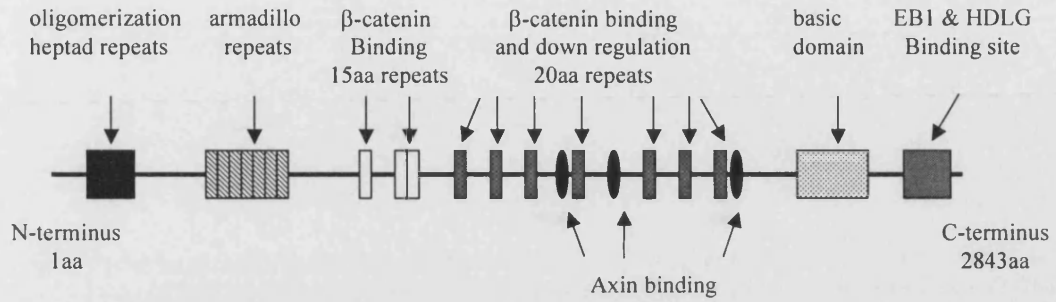


Figure. 5.1. Diagram showing the main domains in the Apc1 protein, adapted from Polakis (2000)

APC2 was identified by virtue of its homology with APC1 from a *Drosophila* expressed- sequence tag (EST) database. The APC2 is highly expressed in the central nervous system as well as the spinal cord and foetal brain. The mouse Apc2 contains an open reading frame of 2274aa, whereas the human APC2 spans over 2302aa (van Es *et al.* 1999). APC1 and APC2 show high similarity in the amino-terminal end of the protein, mainly in the armadillo repeat region; this homology is lost in the carboxy-terminal end of the protein. Although most of the specific domains of the protein are present in APC2, the 15aa repeats seem to be absent in the protein. Another interesting characteristic is that APC2 contains only six of seven of the 20aa repeats of APC1.

The SAMP domains specific for axin binding, have been found to interact with conductin *in vitro*; the APC1 contains three SAMP domains located after the third, fourth and seventh 20aa repeats, All of the domains are necessary for this interaction. In contrast, APC2 contains only two SAMP domains placed after the fourth and seventh 20aa repeats.

Although the complete functions of APC2 have not been completely elucidated, studies indicate that both APC proteins have overlapping functions. Studies performed in *Drosophila* reveal that the armadillo repeats in APC2 are sufficient to regulate the Arm protein levels. However, the two APC proteins show different cellular distributions, APC2 is localised in the cortex of the cells, whereas APC1 is mainly localised in cytoplasmic microtubules.

Comparative genomics allows for the identification of homologous genes in higher organisms. We used homologies with the mouse and human *Apc1* and *Apc2* to characterise and describe the *Fugu Apc1* and *Apc2*. We have compared transcriptional organisation, gene structure and amino acid sequence to analyse how well conserved these genes are across species. Furthermore, we looked for conserved protein domains and tissue distribution of the *Fugu Apc1* and *Apc2* genes.

5.2. Materials and methods

5.2.1 Expression analysis of *Fugu Apc1* by RT-PCR

Total RNA from pufferfish wild type adult and 2.2 and 4.6 day embryos were obtained from Dr. Greg Elgar (MRC UK HGMP Resource Centre). The primers designed for *frApc1* expression are: forward primer 5'- aggagagggatgccgactgt -3' and reverse primer 5'- ggttctcctgccaactccaa -3'. Internal primers for the β-actin gene were designed to confirm that there was no contamination with genomic DNA. The β-actin primers were: forward primer 5'-tacagactacctcatgaagatcc-3' and the reverse primer 5'-gaggccag gatggagcctcc-3'.

For the *frApc1*, the PCR was performed in a 20μl reaction containing 1μl cDNA product from the reverse transcription, 10mM dNTPs, 25mM primers, 2 units Taq polymerase, 1.5mM $MgCl_2$ and 1X Buffer (BioLine). A PTC-225 Peltier Thermal Cycler (MJ Research) was used with the following program: 95^0C, 2 min; 35 cycles at 95^0C for 30s; 65^0C for 30s; 72^0C for 40s; final extension was at 72^0C for 5min. The PCR product was separated in a 1% agarose gel. The PCR product size for the *FrApc1* was 800bp and the PCR product size for the *β-actin* was 500bp.

5.2.2. Expression analysis of *Fugu Apc2* by RT-PCR

Total RNA from pufferfish wild type adult and 2.2 and 4.6 day embryos were obtained from Dr. Greg Elgar (MRC UK HGMP Resource Centre). The primers used for *frApc2* expression are: forward primer 5'-ggctgctgtccattcttacta-3' and reverse primer 5'-aacaaggctggacacatttcg-3'. Internal primers for the β-actin gene were designed to confirm that there was no contamination with genomic DNA. The *β-actin* primers were: forward primer 5'-tacagactacctcatgaagatcc-3' and the reverse primer 5'-gaggccag gatggagcctcc-3'.

For the *frApc2*, the PCR was performed in a 20µl reaction containing 1µl cDNA product from the reverse transcription, 10mM dNTPs, 25mM primers, 2 units Taq polymerase, 1.5mM $MgCl_2$ and 1X Buffer (BioLine). A PTC-225 Peltier Thermal Cycler (MJ Research) was used with the following program: 95^{0}C, 2 min; 35 cycles at 95^{0}C for 30s; 65^{0}C for 30s; 72^{0}C for 40s; final extension was at 72^{0}C for 5min. The PCR product was separated in a 1% agarose gel. The PCR product size for the *FrApc2* was 900bp and the PCR product size for the *β-actin* was 500bp.

5.3. Results

5.3.1. Identification of Fugu genes

The *Fugu Apc* genes were identified using the *Fugu* genomics project database. This work was also carried out after the completion of the *Fugu* genome-sequencing project. The mouse and human *Apc* genes, the transcriptional organisation of the genes, gene structure and aa sequence were taken from the Ensembl database. And in the same way mentioned in chapter 3, we used the information provided from the Ensembl database to identify and characterize the *Fugu Apc* genes.

The *Fugu* genes were identified mainly; by gene structure, nucleotide sequence conservation and amino acid sequence conservation. It was not possible to identify the genes by synteny, because it was not conserved. To identify the *Fugu* Apc1, the mouse Apc1 amino acid sequence (NM_031488) was blasted against the *Fugu* genome database. The first hit was the Mayffold 281 (M000281). The mayffold was analysed using the NIX program, which showed the complete predicted *Apc* gene plus the 5' and 3' region of the gene. By comparative analysis using the aa sequence of the human and mouse Apc1 (discussed below), the predicted *Fugu* gene was designated as the *Fugu Apc1* (*frApc1*).

Using the same strategy, the *Fugu Apc2* was identified blasting the mouse Apc2 aa sequence against the *Fugu* genome database. The best hit was the Mayffold 1664 (M001664), which contains the putative *Fugu Apc* gene. Analysis of the predicted gene by NIX showed a truncated gene located at the beginning of the mayffold. This partial gene contains the last 7 coding exons. To identify the first exons, a blast of the first 7 coding exons against the *Fugu* scaffold database hit the scaffold 11727; which contains the first and second coding exons of the gene

(described below). By comparative analysis, the gene contained in the M001664 and scaffold 11727 was assigned as *Fugu Apc2* (*frApc2*).

5.3.2. Transcriptional organisation of Fugu Apc genes

5.3.2.1. Transcriptional organisation of *Fugu Apc1*

The mouse and human *Apc1* genes have been sequenced and characterized previously. The chromosomal locations of the genes provided by the Ensembl database were used to determine the intergenic distances between *Fugu Apc1* and its neighbouring genes. For the *frApc1*, the intergenic distances were calculated using the locations predicted by the NIX program in the Mayffold.

The human and mouse *Apc1* genes are located on the chromosomes 5 and 18 respectively. Synteny is well conserved between these two species; the nearest gene to the 5' end is the *EPB4.1L4A* for the human and the *Epb4.1l4a* for the mouse. The second nearest gene to this end is the neural protein *Csorf13* gene in the human and related lipid transferred protein 4 *Sartd4* in mouse. At the 3' end of the *APC1*, the nearest gene in human is the signal recognition particle *SRP19*; in the mouse it is the polyposis locus protein 1 *Dp1*. The second nearest gene at this end in the human is the *DP1* gene and in the mouse the polycystic kidney 2-like protein *PKD212* gene (Figure 5.2).

The *Fugu Apc1* is located in the Mayffold 281 in the *Fugu* database. This mayffold contains the complete gene and its flanking neighbouring genes. The immediate 5' gene to the *Fugu Apc1* is the *Fukutin* gene (*FR*) located 4.12Kb from the gene; the second nearest gene at this end is the glyceraldehyde-3-phosphatase dehydrogenase *GPD* gene, placed 13Kb away from the *frApc1*. At the 3' end, the closest gene is the alpha fucosyltransferase *αFT* gene, at 2.49kb; the second gene is a glycoprotein, the low-density lipoprotein receptor *Megalin*, situated 39.8Kb from the *frApc1* (Figure 5.2).

Synteny of *Fugu Apc1* seems broken when compared to the transcriptional orientation of the human and mouse; none of the neighbouring genes is conserved in *Fugu*. The intergenic distances in *Fugu* have been reduced almost 52 times compared to the mouse and 77 times compared to the human in the 5' region. The 3' region is also reduced 10.5 times compared to the mouse and approximately 6 times compared with the human.

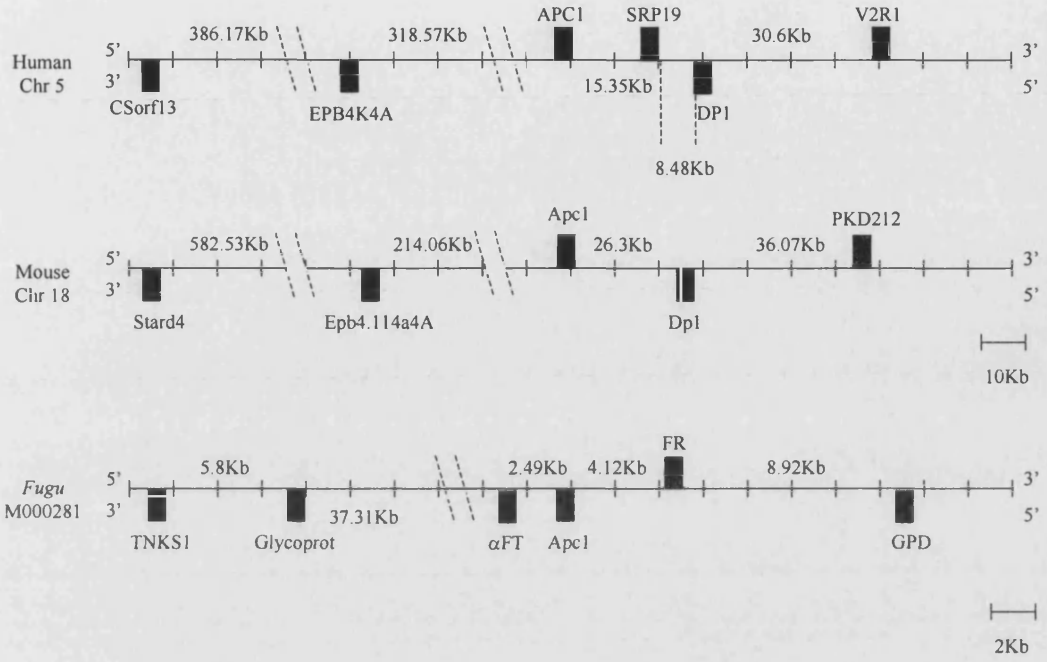


Figure. 5.2. Transcriptional organisation of the *Apc1* gene in human mouse and *Fugu*. The *Apc1* genomic regions from human and mouse were extracted from Ensembl. The *Fugu Apc1* genomic region (M000281) was extracted from the *Fugu* database. Black boxes indicate the position of each gene through the chromosome; genes located above the line are transcribed 5'→3'; genes located under the line are positioned in the reverse DNA strand and are transcribed in opposite orientation. The scale bar for the human and mouse is 10Kb.

5.3.2.2. Transcriptional organisation of *Fugu Apc2*

The human and mouse *Apc2* genes have also been previously characterized. The human gene is located on the chromosome 19 and the mouse gene on the chromosome 10; genomic organisation is well conserved between the mouse and human *Apc2*. The proximal 5' gene in the two species is the ribosomal protein S15 (*Rsp15*), the second proximal gene for both species is the Daz associated protein 1 (*Dazap1*). At the 3' end, the gene immediately next to the human *Apc2* is *C19orf25*, which is a predicted gene without an assigned function or name; the second neighbouring gene is convertase subtisilin kekintype 4 gene (*PCSK4*). In the mouse, the nearest 3' gene to *Apc2* is a predicted gene named Rik; the second proximal gene is, as in the human, the *Pcsk4* gene (Figure 5.3).

The *Fugu Apc2* is placed in the mayffold 1664; seven exons corresponding to the last 7 coding exons of the mouse and human, exons 8 to 14, are located at the beginning of the mayffold; no 5' neighbouring genes are located at this end. The blast of the amino acid region corresponding to the first 7 exons, did not hit any mayffold in the *Fugu* database, however, the blast of the same region against the *Fugu* scaffolds database showed that the first two coding exons of the gene are present in the scaffold 11727 (discussed later); no neighbouring genes were predicted for this scaffold using the NIX program. The two proximal neighbouring genes predicted in the 3' end of the *frApc2* are the ribonucleoprotein *RP* gene, situated 4.6Kb from the gene, and a predicted gene located 13Kb away from the gene (Figure 5.3).

Although synteny is well conserved between mouse and human *Apc2*, the synteny conservation is broken in *frApc2*; even the intergenic distances are larger in *Fugu* than in mouse and human.

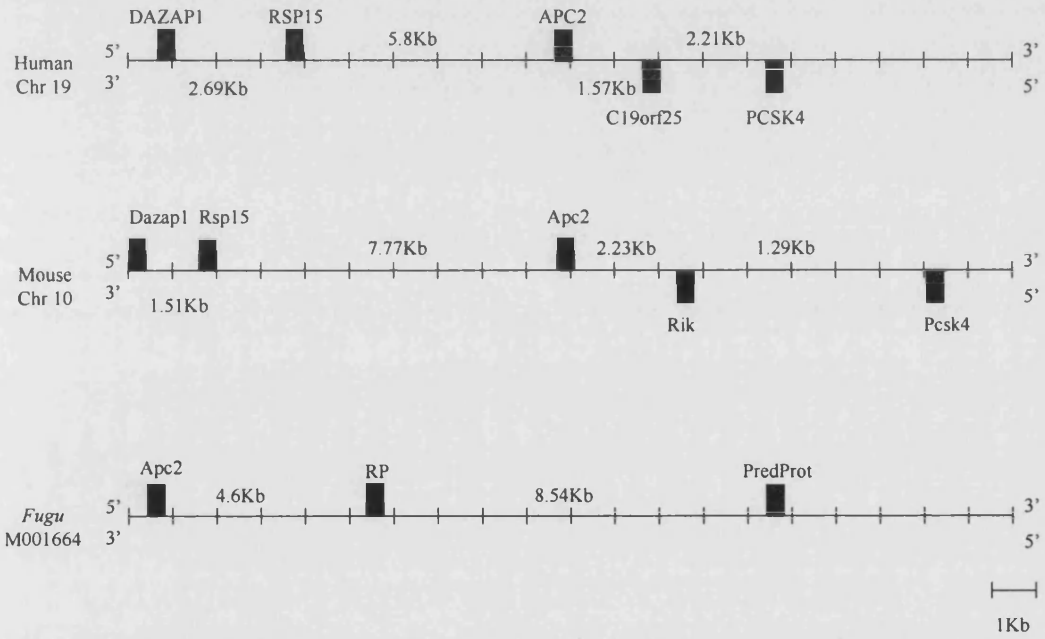


Figure. 5.3. Transcriptional organisation of the *Apc2* gene in human mouse and *Fugu*. The *Apc2* genomic regions of human and mouse were extracted from Ensembl. The *Fugu Apc2* genomic region (M0001664) was extracted from the *Fugu* database. Black boxes indicate the position of each gene through the chromosome; genes located above the line are transcribed 5'→3'; genes located under the line are positioned in the reverse DNA strand and are transcribed in opposite orientation. The scale bar is 1Kb.

5.3.3. Gene organisation of Fugu Apc genes

5.3.3.1. Gene structure of *Fugu Apc1*

Comparative analyses using the coding sequences of human and mouse *Apc1*, and analysis of the splicing acceptor donors, show that the *Fugu Apc1* contains 15 coding exons and 14 introns placed in between each exon, extending over 6768bp, almost 2Kb shorter than the mouse and human gene. The structure of the gene is preserved among the three species; the length of the first 14 exons is well conserved, except for the 6th and 8th exons of *Fugu*, which are shorter than the human and mouse exons. The sixth *Fugu* exon is 20bp shorter, whereas the eighth exon is 56bp shorter; however, the exon-intron phases are preserved (Figure 5.4). The 15th exon, which is the same length in mouse and human, is approximately 2Kb shorter in *Fugu*.

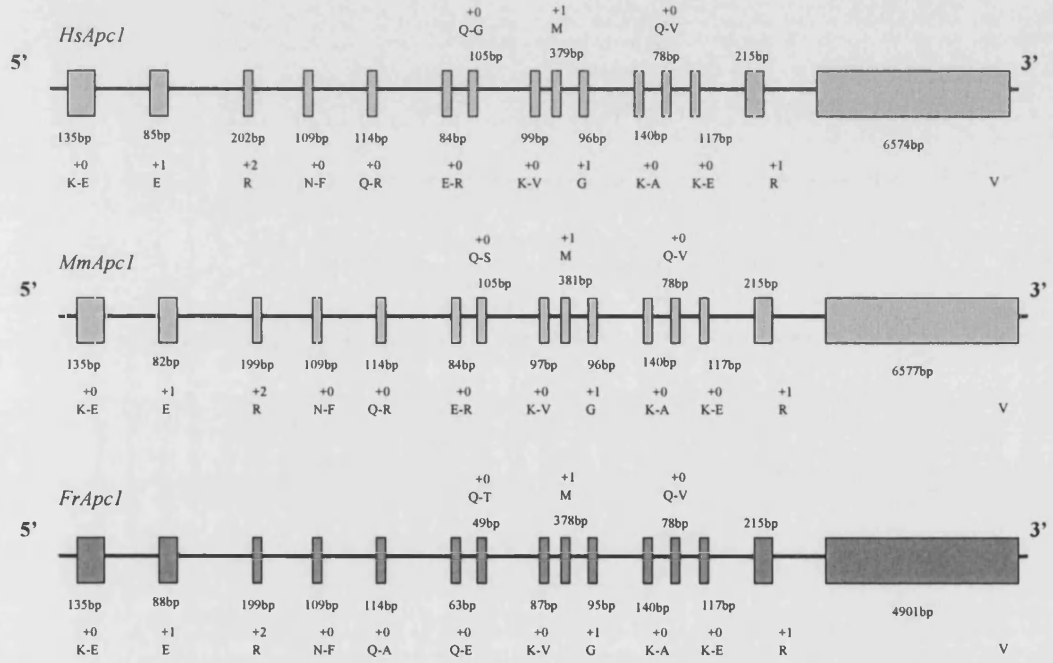


Figure. 5.4. Gene structure of the human, mouse and *Fugu Apc1*. The *Apc1* gene is organized in 15 coding exons and 14 introns. Boxes indicate the exons. The exon length, intron-exon phase and the amino acid of each boundary is indicated for each exon.

The nucleotide coding sequence shows a high level of conservation among the three species. Each coding exon was aligned individually using the ClustalW program; percentage of conserved nucleotides was estimated for each exon using the

Fugu sequence as a base of comparison. Table 5.1 shows the conserved nucleotides of each exon among the three species. Exons 1, 4, 9, 10, 11, 12 and 13 are the best-preserved exons of the gene with more than 70% conservation, whereas exons 3, 7 and 15 show low level of conservation. Exon 3, which shares the same length as the human and mouse, is 43.21% conserved. The low percentage of conservation in the exon 7 is due to the length of the exon, which is shorter in *Fugu*. The exon 15, the largest exon of the gene, is just 51.41% conserved among the three species (Appendix section 3A).

Exon	Num. of nt conserved	Percentage of conservation
1	104/135	77.03%
2	51/88	57.95%
3	86/199	43.21%
4	81/109	74.31%
5	68/114	59.64%
6	34/63	53.96%
7	19/49	38.77%
8	52/87	59.77%
9	271/378	71.69%
10	78/96	81.25%
11	100/140	71.42%
12	61/78	78.20%
13	83/117	70.94%
14	147/215	68.37%
15	2520/4901	51.41%

Table 5.1. Nucleotide conservation of *Apc1* exons among mouse, human and *Fugu*. The deduced *Fugu* exons were used for comparison. The percentage of conservation was calculated based on the conserved number of nucleotides among the three species and using the *Fugu* sequence for comparison.

5.3.3.2. Gene structure of the *Fugu Apc2*

The *Apc2* gene structure of the human and mouse consists of 14 coding exons and 13 introns. As in the Apc1 gene, the last exon is the largest of the gene. The human *Apc2* extends over 6909 nucleotides, whereas the mouse *Apc2* is 6831nt in length.

The *frApc2* structure was obtained using the intron-exon boundaries of each exon and the alignment of the mouse and human exons against the putative *Fugu*

exons. For the present exons of the *frApc2*, the intron-exon faces are well conserved. The putative *frApc2* gene extends over 6800nt; 9 coding exons of the 14 exons were identified; exons 1, 2, 8, 9, 10, 13 and 14 are present in the gene; exons 11 and 12 are fused to form one exon which is 195bp in length, the same length than the mouse and human exon 11 and 12 together. The *Fugu* exon 8 is 400bp in length, 9 and 24bp larger than the human and mouse 8th exon respectively. The *frApc2* exon 14 extends over 5543bp, 484bp larger than the human exon and 550bp larger than the mouseexon; the *Fugu* exon 14 contains sequences of 10 to 20bp, which are not present either in human and mouse (Figure 5.5). We were unable to identify the exons 3 to 7 of the gene in the *Fugu* database.

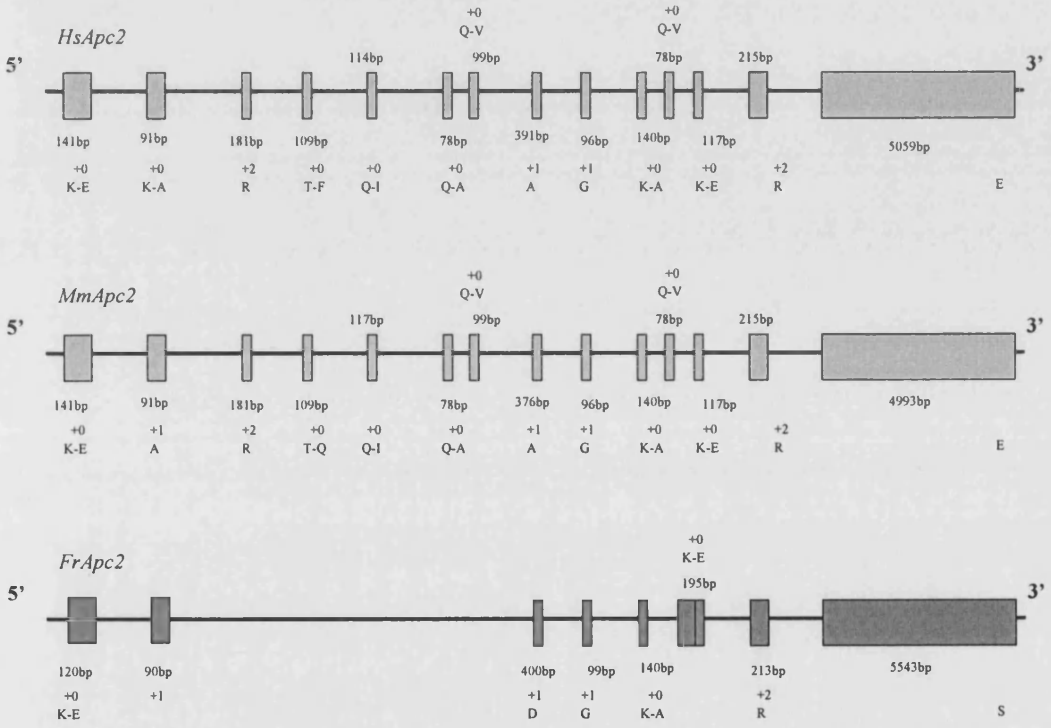


Figure. 5.5. Gene structure of the human, mouse and *Fugu Apc2*. The *Apc2* gene is organized in 14 coding exons and 13 introns. Boxes indicate the exons. The exon length, intron-exon phase and the amino acid of each boundary is indicated for each exon.

Nucleotide comparative analysis of t *frApc2* showed that the existing exons of the gene have a high degree of conservation with the human and mouse *Apc2* exons. Each coding exon was aligned individually; percentage of conservation was estimate based on the *Fugu* sequence. The exons 1, 2, 9, 10, 11, 12 and 13 show more than

60% conservation. Exons 8 and 14 show 55.75 and 39.97% conservation respectively, this low conservation may be due to the difference in the length of the exons (Table 5.2 and Appendix section 3B).

Exon	Num. of nt conserved	Percentage of conservation
1	78/120	65.0%
2	59/90	65.55%
3	-	-
4	-	-
5	-	-
6	-	-
7	-	-
8	223/400	55.75%
9	64/99	64.64%
10	90/140	64.28%
11	51/78	65.38%
12	81/117	69.23%
13	161/213	75.58%
14	2142/5359	39.97%

Table 5.2. Nucleotide conservation of the *Apc2* exons among mouse, human and *Fugu*. Alignment of the deduced *Fugu* exons with the mouse and human *Apc2* exons, the percentage of conservation were calculated based on the conserved number of nucleotides among the three species. The *Fugu* sequence was used to estimate the percentages.

5.3.4. Expression analysis of Fugu Apc by RT-PCR

5.3.4.1. Expression of *frApc1* gene in *Fugu* adult tissues

To investigate the *frApc1* expression in *Fugu*, RT- PCR was carried out in wild type embryos and adult tissues from *Fugu*. Expression of *frApc1* was detected in gut, spleen, gonads, kidney, eye, spinal cord, gill and heart in the adult *Fugu*, expression was also present in the *Fugu* embryo at 53.3hpf. an early somites stage and 111.5hpf. In the first stage, the developing organs are starting to form and at 111.5 hpf., all the organs have gone to completion, although the *Fugu* embryo has not hatched yet (Figure 5.6).

These data indicate that the expression of *frApc1* is present in the majority of adult organs in *Fugu*. Furthermore, expression is also present in a very early stage of

development, which is the somites stage. This expression continues during development through the adult organism. Expression of *frApc1* seems to be unvarying during development and in the adult tissues except for the kidney, which shows a slightly reduced presence of transcript. In the case of the gut tissue, the presence of *frApc1* expression does not indicate if the transcript is present in all the intestinal tissue, in an anterior or posterior region, or in a specific group of cells.

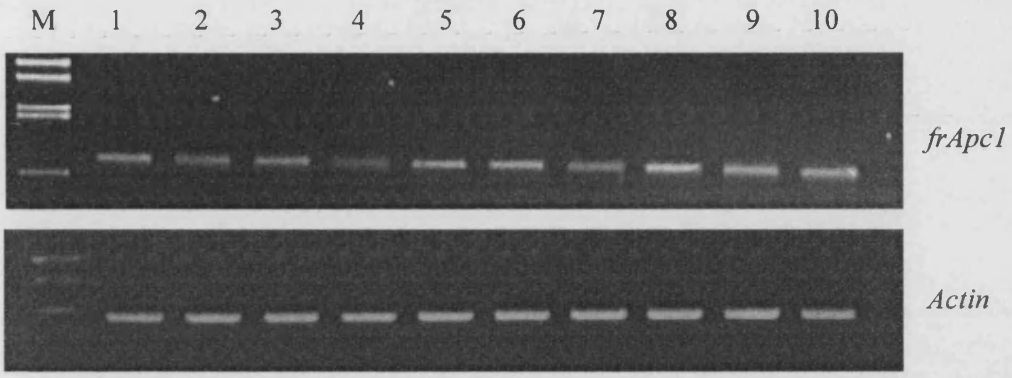


Figure. 5.6. *Fugu rubripes Apc1* mRNA in adult tissues. The *Fugu Apc1* mRNA-specific primers amplified a PCR product from total RNA extracted from the gut (1), spleen (2), gonads (3), kidney (4), eye (5), spinal cord (6), gill (7) heart (8), 53.3hpf embryo (9) and 111.5hpf embryo (10). A RT-PCR product corresponding to the constitutively expressed *Actin* gene was used as internal control.

5.3.4.2. Expression of *frApc2* gene in *Fugu* adult tissues

To look for the *frApc2* expression, RT- PCR was also performed in wild type embryos and adult tissues from *Fugu*. Expression of *frApc2* was detected in heart, gill and gonads in adult tissues; however, expression was not observed in spleen, spinal cord, eye, kidney and gut. Expression of *frApc2* was also identified in the developing *Fugu* embryo at 53.3 hpf. that is, the 5-6 somites stage and at 111.5 hpf. a later developmental stage (Figure 5.7).

In contrast with the expression of *frApc1*, *frApc2* expression seems to be restricted to specific organs in the adult *Fugu*. The *frApc2* transcripts were present in both of the developing stages of the embryo. As in the *frApc1*, *frApc2* expression seems identical during development and in the adult tissues.

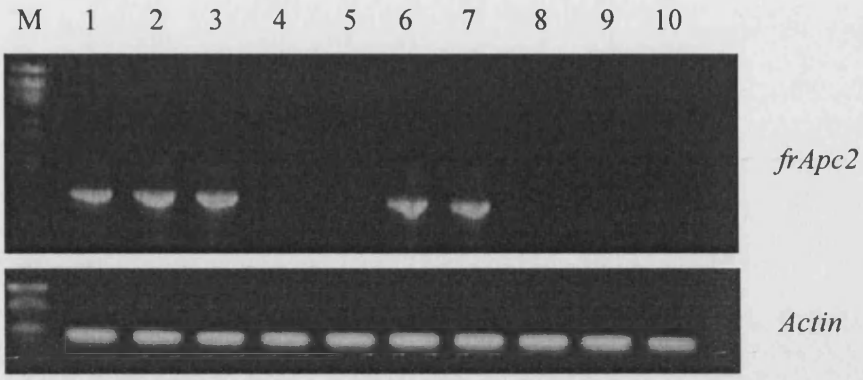


Figure 5.7. *Fugu rubripes Apc2* mRNA in adult tissues. The Fugu Apc2 mRNA-specific primers amplified a PCR product from total RNA extracted from the gonads (1), heart (2), gill (3), spleen (4), spinal cord (5), kidney (6), 111.5hpf embryo (7) 53.3hpf embryo (8), eye (9) and gut (10). A RT-PCR product corresponding to the constitutively expressed Actin gene was used as internal control.

Previous studies have identified the expression of the murine *Apc1* and *Apc2* in a repertoire of different adult tissues and during developmental stages. The *Apc1* expression has a similar tissue distribution in mouse and *Fugu*; the gene is also expressed during developmental stages in both species (Table 5.3).

The mouse and *Fugu Apc2* expression have some similarities and some differences. In mouse and *Fugu* the gene is expressed in the developing embryo. In the adult, the gene is expressed in the gonads in both species. In *Fugu* expression is also seen in heart and spleen. In mouse, the gene is not expressed in heart and spleen expression is seen in spinal cord and eye. No expression is seen in both species in the kidney and gut (Table 5.3).

Spp.	Embryo stage and adult tissue							
	Embryo	Gonads	Heart	Spleen	Spinal cord	Kidney	Eye	Gut
frApc1	+ (E4.6, E2.2)	+	+	+	+	+	+	+
mmApc1	+ (E15, 16)	+	+	+	+	+	+	+
frApc2	+ (E4.6, E2.2)	+	+	+	-	-	-	-
mmApc2	+ (E10,11, 13,14,15, 18)	+	-	-	+	-	+	-

Table 5.3. Comparison of *Apc1* and *Apc2* gene expression in *Fugu* and mouse. Expression of the murine *Apc2* in the embryo at E10, 11, 13 and 15 was characterized by *in situ* hybridisation, the E14 and E18 stages were done by northen blot according with the MGI database (www.informatics.jax.org).

5.3.5. Analysis of the Fugu Apc protein sequence

5.3.5.1. Amino Acid sequence of the *Fugu* Apc1

A comparison of the deduced amino acid sequence of the *frApc1* with the human and mouse orthologues indicates that the translated exons are well conserved among the three species. The coding sequence of the *Fugu* Apc1 extends over 2199aa; it is 642aa shorter than the human protein. The table 5.4 shows the amino acids conserved in each exon and their percentage of conservation; many of the exons show more than 90% conservation when compared with the mouse and human exons. The exons 3 and 7 are the less conserved exons with 68.18% and 53.33% conservation respectively, the exon 15, which codes for more than half of the protein and contains many of the main domains of the protein is 78.92% conserved.

Exon	Num. of aa conserved	Percentage of conservation
1	45/45	100%
2	28/31	90.32%
3	45/66	68.18%
4	35/35	100%
5	35/38	92.10%
6	17/21	80.95%
7	8/15	53.33%
8	28/29	96.55%
9	122/126	96.82%
10	32/34	94.11%
11	45/45	100%
12	26/26	100%
13	36/39	92.30%
14	71/73	97.26%
15	1243/1575	78.92%

Table 5.4. Amino acid conservation of Apc1. Alignment of the aa sequence of each exon was done to estimate the number of aa conserved and percentage of conservation among human, mouse and *Fugu*. The *Fugu* aa sequence was used to estimate the level of conservation.

Previous studies have identified and characterised specific domains in the human and mouse Apc1; the mouse Apc1 was used to identify those specific domains in the *Fugu* Apc1 protein (Fearnhead *et al.* 2001). The table 5.5 shows the conservation of each domain among human, mouse and *Fugu*.

The seven main domains of the protein are present and well conserved in the frApc1, the 7 armadillo repeats, which are located between the 453-767aa, were analysed as one domain. The 15aa repeat located between the 1020- 1170aa were also analysed as one domain. And the 20aa repeats which are series of seven repeats, were characterized by their specific core sequence TPXXFSXXXSL (Groden *et al.* 1991).

Amino acid	Domain	Num. of aa conserved	Percentage of conservation
6- 57	Oligomerisation domain	51/51	100%
453- 767	Armadillo region	305/314	97.13%
1020- 1170	15 aa repeat	53/150	35.33%
1265- 2035	20 aa repeat	13/14	92.85%
		14/14	100%
		15/16	93.75%
		11/11	100%
		13/14	92.85%
		14/14	100%
		14/14	100%
2200- 2400	Basic Domain	86/200	43.0%
2559- 2771	EB1 binding site	130/212	61.33%
2772- 2843	HDLG binding site	49/72	68.05%

Table 5.5. Specific domain conservation of Apc1. Specific domains were delimited using the described domains in the mouse Apc1. The number of aa and percentages are the conservation among human, mouse and *Fugu* domains.

The 3 SAMP repeats are also conserved in the frApc1, they are placed after the 3rd, 4th and 7th 20aa repeats. The first repeat is 80% (24/30aa) conserved with the mouse and human. The second repeat is 14aa in length; it is 15aa shorter than the mouse and human and 12 of the 14aa are conserved (85.71%). The third repeat is 88.88% conserved (24/27aa) with the mouse and human repeat (Figure 5.8)

```
hsE1 MAAASYDQLLKQVEALKMENSNLRQELEDNSNHLTKLETEASNMK 45
mmE1 MAAASYDQLLKQVEALKMENSNLRQELEDNSNHLTKLETEASNMK 45
frE1 MAAASYDQLLRQVEVLKMENSNLRQELQDNSNHLTKLETEASNMK 45
     **********:***.***********:*****************

hsE2 EVLKQLQGSIEDEA-MASSGQIDLLERLKEL 30
mmE2 EVLKQLQGSIEDET-MTS-GQIDLLERLKEF 29
frE2 EVLKQLQGTIEEESGEASGSQLELIERLKEM 31
     ********:**:*:  :* .*::*:*****:

hsE3 NLDSSNFPGVKLRSKMSLRSYGSREGSVSSRSGECSPVPMGSFPRRGFVNGSRESTG-YLEELEKERS 67
mmE3 NLDS-NFPGVKLRSKMSLRSYGSREGSVSSRSGECSPVPMGSFPRRTFVNGSRESTG-YLEELEKERS 66
frE3 SLESAGFK-HRTRPPMPTSSPSPASGSGAPGAAGGGPQASAAFGRRGMPTVGRESHDRCLEELEKERS 66
     .*:* .*   : *. *.  * .. .** :. :.  .* . .:* ** : . .*** .  *********

hsE4 LLLADLDKEEKEKDWYYAQLQNLTKRIDSLPLTEN 35
mmE4 LLLADLDKEEKEKDWYYAQLQNLTKRIDSLPLTEN 35
frE4 LLLAELEKEEKEKDWYYAQLQNLTKRIDSLPLTEN 35
     ****:*:****************************

hsE5 FSLQTDMTRRQLEYEARQIRVAMEEQLGTCQDMEKRAQ 38
mmE5 FSLQTDMTRRQLEYEARQIRAAMEEQLGTCQDMEKRAQ 38
frE5 FTLQTDRSRLQLEFEARQIRSAMEEQLGSCQEMERRAQ 38
     *:**** :* ***:****** *******:**:**:***

hsE6 RRIARIQQIEKDILRIRQLLQSQATEAE 28
mmE6 RRIARIQQIEKDILRVRQLLQSQAAEAE 28
frE6 ARVSRIQQIEKDILRLGAHLQ------- 21
      *::***********:   **

hsE7 RSSQNKHETGSHDAERQNEGQGVGEINMATSGNGQ 35
mmE7 RSSQSRHDAASHEAGRQHEGHGVAESNTAASSSGQ 35
frE7 ------------EVQALGDSSGLAAAQ-------- 15
                 :.    :. *:.  :
```

```
hsE8 GSTTRMDHETASVLSSSSTHSAPRRLTSHLGTK 33
mmE8 SPATRVDHETASVLSSSGTHSAPRRLTSHLGTK 33
frE8 TASSRLDHEP----TSEASYSVPRRITNHLGTK 29
      .::*:***.    :*..::*.***:*.*****

hsE9 VEMVYSLLSMLGTHDKDDMSRTLLAMSSSQDSCISMRQSGCLPLLIQLLHGNDKDSVLLG 60
mmE9 VEMVYSLLSMLGTHDKDDMSRTLLAMSSSQDSCISMRQSGCLPLLIQLLHGNDKDSVLLG 60
frE9 VEMVYSLLSMLGTHDKDDMSRTLLAMSSSQDSCIAMRQSGCLPLLIQLLHGNDKDSILLG 60
     ************************************:**********************:***

hsE9 NSRGSKEARARASAALHNIIHSQPDDKRGRREIRVLHLLEQIRAYCETCWEWQEAHEPGMDQDKNPMP 128
mmE9 NSRGSKEARARASAALHNIIHSQPDDKRGRREIRVLHLLEQIRAYCETCWEWQEAHEQGMDQDKNPMP 128
frE9 NSRGSKEARARASAALHNIVHSQPDDKRGRREIRVLHLLEQERHYCEACWSWQENHERGIDQEDNP-- 126
     ******************:********************** * ***:**.*** ** *:**:.**

hsE10 --APVEHQICPAVCVLMKLSFDEEHRHAMNELGG 32
mmE10 --APVEHQICPAVCVLMKLSFDEEHRHAMNELGG 32
frE10 MPSPVEHQICPAVCVLMKLSFDEEHRHAMNELGG 34
        :*******************************

hsE11 LQAIAELLQVDCEMYGLTNDHYSITLRRYAGMALTNLTFGDVANK 45
mmE11 LQAIAELLQVDCEMYGLTNDHYSVTLRRYAGMALTNLTFGDVANK 45
frE11 LQAVAELLQVDCEMFGLTSDHYSITLRRYAGMALTNLTFGDVANK 45
      ***:**********:***.****:**********************

hsE12 ATLCSMKGCMRALVAQLKSESEDLQQ 26
mmE12 ATLCSMKGCMRALVAQLKSESEDLQQ 26
frE12 ATLCSMKGCMRAMVAQLKSDSEDLQQ 26
      ************:******:******

hsE13 VIASVLRNLSWRADVNSKKTLREVGSVKALMECALEVKK 39
mmE13 VIASVLRNLSWRADVNSKKTLREVGSVKALMECALEVKK 39
frE13 VIASVLRNLSWRADVNSKKTLREVGSVRALTGCALVVQK 39
      ***************************:**  *** *:*
```

```
hsE14 ESTLKSVLSALWNLSAHCTENKADICAVDGALAFLVGTLTYRSQTNTLAIIESGGGILRNVSSLIATNEDHRQ 73
mmE14 ESTLKSVLSALWNLSAHCTENKADICAVDGALAFLVGTLTYRSQTNTLAIIESGGGILRNVSSLIATNEDHRQ 73
frE14 ESTLKSVLSALWNLSAHCTENKADICAVEGALAFLVGTLTHCSHTNTLAIIESGGGILRNVSSLIATNEAHRQ 73
      ****************************:***********: *:************************ ***

hsE15 -ILRENNCLQTLLQHLKSHSLTIVSNACGTLWNLSARNPKDQEALWDMGAVSMLKNLIHS 59
mmE15 -ILRENNCLQTLLQHLKSHSLTIVSNACGTLWNLSARNPKDQEALWDMGAVSMLKNLIHS 59
frE15 QTLREQGCLPTLLQHLKSHSLTIVSNACGTLWNLSARDAKDQETLWELGAVGMLRNLIHS 60
        ***:.** ***************************:.****:**::***.**:*****

hsE15 KHKMIAMGSAAALRNLMANRPAKYKDANIMSPGSSLPSLHVRKQKALEAELDAQHLSETF 119
mmE15 KHKMIAMGSAAALRNLMANRPAKYKDANIMSPGSSLPSLHVRKQKALEAELDAQHLSETF 119
frE15 RHKMIAMGSAAALRNLMANRPARYKDASVVSPGAGAPSLHVRKQKALFEELDAQQLSETF 120
      :**********************:****.::***:. ***********  *****:*****

hsE15 DNIDNLSPKASHRSKQRHKQSLYGDYVFDTNRHDDNRSDNFNTGNMTVLSPYLNTTVLPS 179
mmE15 DNIDNLSPKASHRSKQRHKQNLYGDYAFDANRHDDSRSDNFNTGNMTVLSPYLNTTVLPS 179
frE15 DNIDNLSPKTAHRK----------------------GRGCNSASGTASTARPYTNTPVLSS 159
      *********::**...  ...  ..   .:.  ...*. *  :*. :.  ** **.**.*

hsE15 SSSSRGSLDSSRSEKDRSLERERGIGLGNYHPATENPGTSSKRGLQISTTAAQIAKVMEE 239
mmE15 SSSSRGSLDSSRSEKDRSLERERGIGLSAYHPTTENAGTSSKRGLQITTTAAQIAKVMEE 239
frE15 PKNGDG------------------------------------------------------ 165
       ... *: .:: :... : . . . . .   .::....:::. . . :::::. :.  ..

hsE15 VSAIHTSQEDRSSGSTTELHCVTDERNALRRSSAAHTHSNTYNFTKSENSNRTCSMPYAK 299
mmE15 VSAIHTSQDDRSSASTTEFHCVADDRSAARRSSASHTHSNTYNFTKSENSNRTCSMPYAK 299
frE15 ------------------------------------------------------------
       ::  ::... ::.:::.  . :.. .:   :::: : :.: . :.:..:. :.: . :.

hsE15 LEYKRSSNDSLNSVSSSDGYGKRGQMKPSIESYSEDDES---KFCSYGQYPADLAHKIHS 356
mmE15 VEYKRSSNDSLNSVTSSDGYGKRGQMKPSVESYSEDDES---KFCSYGQYPADLAHKIHS 356
frE15 -----ASSDSLNSVTSADGYGNRGKNKPSTEPFYSSDESGANKCCVYRKYPADLAHKIRS 220
       . . :*.******:*:****:**: *** *.: ..***   * * * :*********:*
```


hsE15 ANHMDDNDG-ELDTPINYSLKYSDEQLNSGRQSPSQNERWARPKHIIEDEIKQSEQRQSR 415
mmE15 ANHMDDNDG-ELDTPINYSLKYSDEQLNSGRQSPSQNERWARPKHVIEDEIKQNEQRQAR 415
frE15 ANHMADDDGAELDTPINYSLKYSDEQLNSGRQSPSHRVGMD------SDDDDEEDGRLRR 274
hsE15 NQSTTYPVYTESTDDKHLKFQPHFGQQECVSPYRSRGANGSETNRVGSNHGINQNVSQSL 475
mmE15 SQNTSYPVYSENTDDKHLKFQPHFGQQECVSPYRSRGTSGSETNRMGSSHAINQNVNQSL 475
frE15 RNDGSDSVSSGS-------------RIISVPPPRYVVTAAAAN---------------- 304
hsE15 CQEDDYEDDKPTNYSERYSEEEQHEEEE-RPTNYSIKYNEEKRHVDQPIDYSLKYATDIP 534
mmE15 CQEDDYEDDKPTNYSERYSEEEQHEEEEERPTNYSIKYNEEKHHVDQPIDYSLKYATDIS 535
frE15 -----YGGDPAG---------------------------------EQPIDYSLKYGSDG- 325
hsE15 SSQKQSFSFSKSSSGQSSKTEHMSSSSENTSTPSSNAKRQNQLHPSSAQSRSGQPQKAAT 594
mmE15 SSQKPSFSFSKNSSAQSTKPEHLSPSSENTAVPPSNAKRQNQLRPSSAQ-RNGQTQKGTT 594
frE15 -AHKPLF-----------KPEESAASSVKTPTPSSSAKLR--------------PPAPAN 359
hsE15 CKVSSINQETIQTYCVEDTPICFSRCSSLSSLSS-AEDEIGCNQTTQEADSAN-TLQIAE 652
mmE15 CKVPSINQETIQTYCVEDTPICFSRCSSLSSLSS-ADDEIGCDQTTQEADSAN-TLQTAE 652
frE15 RAVAKANQESTQTYCVEDTPICFSRGSSLSSLSSEEEEEVDVIERRGASGGGNGEYPTVP 419
hsE15 IKEKIGTRSAEDPVSEVPAVSQHPRTKSSRLQGSSLSSESARH-KAVEFSSGAKSPSKSG 711
mmE15 VKENDVTRSAEDPATEVPAVSQNARAKPSRLQASGLSSESTRHNKAVEFSSGAKSPSKSG 712
frE15 VSEKDAREQQQRHQKEAESQTAAVTAPSTRGRRS-----HHHHHHGHHHHHHHHVASSSG 474
hsE15 AQTPKSPPE-HYVQETPLMFSRCTSVSSLDSFESRSIASSVQS-EPCSGMVSGIISPSDL 769
mmE15 AQTPKSPPE-HYVQETPLVFSRCTSVSSLDSFESRSIASSVQS-EPCSGMVSGIISPSDL 770
frE15 ARTPKSPPEPPYAQETPLMFSRCTSVSSLDSFSTSSIASSVRSSEPCSGMPSGVVSPSDL 534

```
hsE15 PDSPGQTMPPSRSKTPPPPPQTAQTKREVPKNKAPTAEKRESGPKQAAVNAAVQRVQVLP 829
mmE15 PDSPGQTMPPSRSKTPPPPPQTVQAKREVPKSKVPAAEKRESGPKQTAVNAAVQRVQVLP 830
frE15 PDSPGQTMPPSRAKTPPLPPTTHKTNKQENTKKKDEDE--------------------S 573
      ************:**** ** * :::::  ..*    *. .:....:: .:: .  .  .

hsE15 DADTLLHFATESTPDGFSCSSSLSALSLDEPFIQKDVELRIMPPVQENDNGNETESEQPK 889
mmE15 DVDTLLHFATESTPDGFSCSSSLSALSLDEPFIQKDVELRIMPPVQENDNGNETESEQPE 890
frE15 SADVLLHFATESTPHGFSRASSLSALSVDEPFITAELKE------EEEDKGSEQRVEEAP 627
      ..*.**********.*** :*******:*****  :::    .. :*:*:*.* . *:.

hsE15 ESNENQEKEAEKTIDSEKDLLDDSDDDD-IEILEECIISAMPTKSSRKAKKPAQTASKLP 948
mmE15 ESNENQDKEVEK-PDSEKDLLDDSDDDD-IEILEECIISAMPTKSSRKAKKLAQTASKLP 948
frE15 -----------------KAVLDESDDDDDIEILEACINMAMP-KSSRKPKKPQQAAP--- 666
      .:........ ..: .:.* :**:***** ***** **  ***:*****.**  *:*.. .

hsE15 PPVARKPSQLPVYKLLPSQNRLQPQKHVSFTPGDDMPRVYCVEGTPINFSTATSLSDLTI 1008
mmE15 PPVARKPSQLPVYKLLPAQNRLQAQKHVSFTPGDDVPRVYCVEGTPINFSTATSLSDLTI 1008
frE15 ----RKASQLPVYKLRVQS---QPRKDVP-PPAEEVPRVYCVEGTPLNFSTATSLSDLTI 718
      .. :**.********   ..  *.:*.*. .*.:::**********:**************

hsE15 ESPPNELAAGEGVRGGAQSGEFEKRDTIPTEGRSTDEAQGGKTSSVTIPELDDNKAEEGD 1068
mmE15 ESPPNELATGDGVRAGIQSGEFEKRDTIPTEGRSTDDAQRGKISSIVTPDLDDNKAEEGD 1068
frE15 DSPPNEEAA--------------------------------AVAAAILPEGPPASTQED- 745
      :***** *:...  .. .:.. .. .: .:.. ::..:. .  ::   *:    .::*..

hsE15 ILAECINSAMPKGKSHKPFRVKKIMDQVQQASASSSAPNKNQLDGKKKKPTSPVKPIPQN 1128
mmE15 ILAECINSAMPKGKSHKPFRVKKIMDQVQQASSTSSGANKNQVDTKKKKPTSPVKPMPQN 1128
frE15 ------RAGHPEGE-----------------NAMTSLPNVSALLCPKPNPGS-------Q 775
        :.. .:. *:*:: ..   ..  .. ..:.: :* .* . :   * :* *. .. ..:

hsE15 TEYRTRVRKNADSKNNLNAERVFSDNKDSKKQNLKNNSKVFNDKLPNNEDRVRGSFAFDS 1188
mmE15 TEYRTRVRKNTDSKVNVNTEETFSDNKDSKKPSLQTNAKAFNEKLPNNEDRVRGSFALDS 1188
frE15 SEWQMTTTTAAKTKP-------------------------------------GFAFDS 796
      :*::  . . :.:* . .:.   :....:.. . ...:.  ... .....   ..**:**
```


hsE15 PHHYTPIEGTPYCFSRNDSLSSIDFDDDDVDLSREKAELRKAKENKESEAKVTSHTELTS 1248
mmE15 PHHYTPIEGTPYCFSRNDSLSSIDFDDDDVDLSREKAELRKGKESKDSEAKVTCRPEPNS 1248
frE15 PRHYTPIEGTPCCFSRNDSLSSIDFDEDDGGDKDEEHKKTKEEDGK---------- 842
hsE15 NQQSANKTQAIAKQPINRGQPKPILQKQSTFPQSSKDIPDRGAATDEKLQNFAIENTPVC 1308
mmE15 SQQAASKSQASIKHPANRAQSKPVLQKQPTFPQSSKDGPDRGAATDEKLQNFAIENTPVC 1308
frE15 ----------KRKQQTAAVFPRTKAN-----ATSSDEKQKFAIEDTPVC 876
hsE15 FSHNSSLSSLSDIDQENNN-KENEP----------IKETEPPDSQGEPSKPQASGYAPK 1356
mmE15 FSRNSSLSSLSDIDQENNNNKESEP----------IKEAEPANSQGEPSKPQASGYAPK 1357
frE15 FSRNSSLSSLSDIDQENNNKEFAPPPPAEQQDGGKAVKSSPPRRAESKPRPPAASGYAPK 936
hsE15 SFHVEDTPVCFSRNSSLSSLSIDSEDDLLQECISSAMPKKKKPSRLKGDNEKHSPRNMGG 1416
mmE15 SFHVEDTPVCFSRNSSLSSLSIDSEDDLLQECISSAMPKKKRPSRLKSESEKQSPRKVGG 1417
frE15 AFHVEDTPVCFSRNSSLSSLSIDSEDDLLQECISSAMPKKKKKAAASATPSTVAAPPAAP 996
hsE15 ILGEDLTLDLKDIQRPDSE--HGLSPDSENFDWKAIQEGANSIVSSLHQ-AAAAACLSRQ 1473
mmE15 ILAEDLTLDLKDLQRPDSE--HAFSPDSENFDWKAIQEGANSIVSSLHQAAAAAACLSRQ 1475
frE15 KAENSILAEEEPPEMPSEVPRSPASPDSESFDWKAIQEGANSIVSSLNAAAAAATSLSRQ 1056
hsE15 ASSDSDSILSLKSGISLGSPFHLTPDQ---------EEKPFTSNKGPRILKPGEKSTLET 1524
mmE15 ASSDSDSILSLKSGISLGSPFHLTPDQ---------EEKPFTSNKGPRILKPGEKSTLEA 1526
frE15 ASSDSDSVLSLKS---VGSPFHLPSANNNAEEDKVDEAEEVAVKRGARILKAGERTTLDA 1113
hsE15 KKIESE--SKGIKGGKKVYKSLITGKVRSNSEISGQMKQPLQANMPSISRGRTMIHIPGV 1582
mmE15 KKIESE--NKGIKGGKKVYKSLITGKIRSNSEISSQMKQPLPTNMPSISRGRTMIHIPGL 1584
frE15 KKEEDEEEAKGVRGGKKVYRSLITGKVRAEE----------AARGR-------- 1149

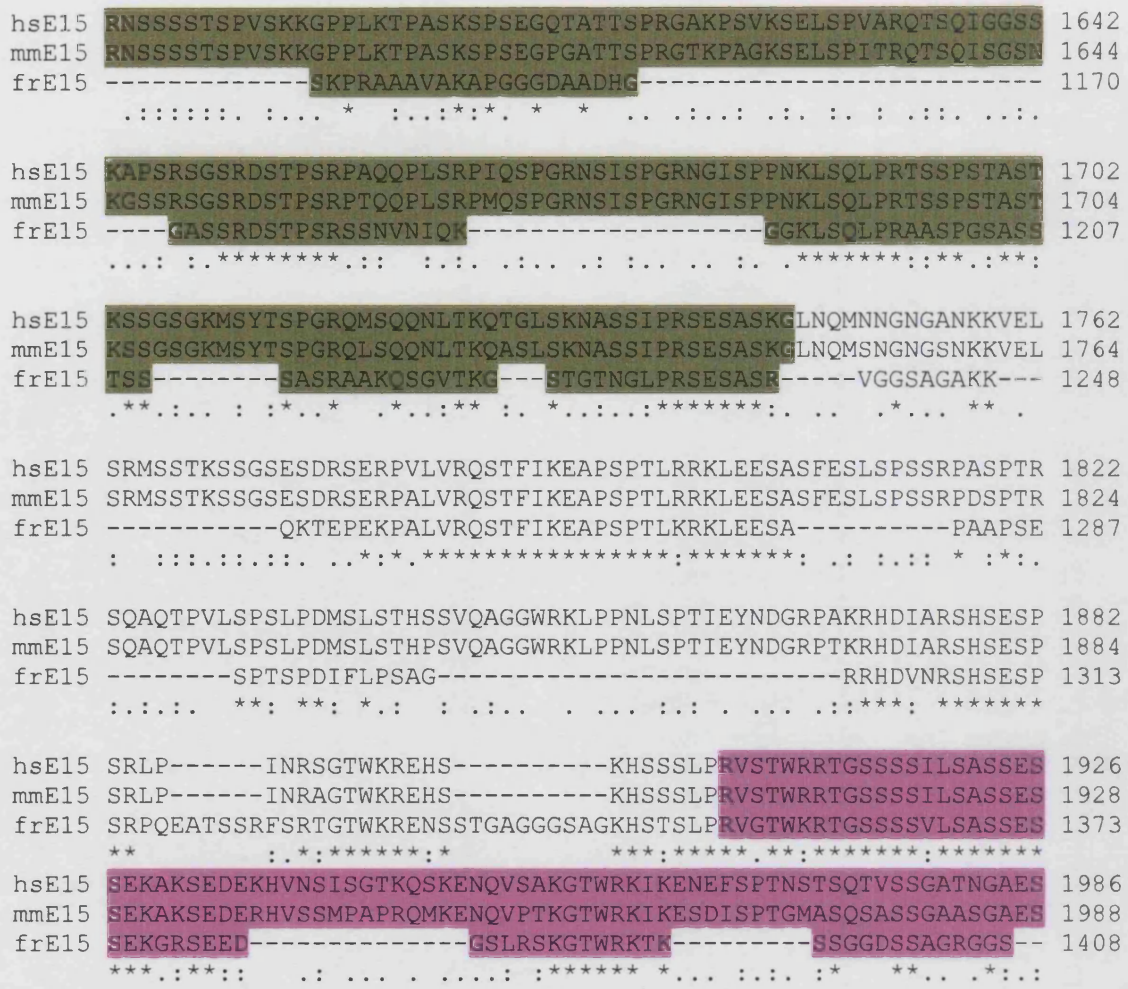

hsE15 RNSSSSTSPVSKKGPPLKTPASKSPSEGQTATTSPRGAKPSVKSELSPVARQTSQIGGSS 1642
mmE15 RNSSSSTSPVSKKGPPLKTPASKSPSEGPGATTSPRGTKPAGKSELSPITRQTSQISGSN 1644
frE15 ----------SKPRAAAVAKAPGGGDAADHG---------- 1170
hsE15 KAPSRSGSRDSTPSRPAQQPLSRPIQSPGRNSISPGRNGISPPNKLSQLPRTSSPSTAST 1702
mmE15 KGSSRSGSRDSTPSRPTQQPLSRPMQSPGRNSISPGRNGISPPNKLSQLPRTSSPSTAST 1704
frE15 ----GASSRDSTPSRSSNVNIQK----------GGKLSQLPRAASPGSASS 1207
hsE15 KSSGSGKMSYTSPGRQMSQQNLTKQTGLSKNASSIPRSESASKGLNQMNNGNGANKKVEL 1762
mmE15 KSSGSGKMSYTSPGRQLSQQNLTKQASLSKNASSIPRSESASKGLNQMSNGNGSNKKVEL 1764
frE15 TSS--------SASRAAKQSGVTKG---STGTNGLPRSESASR-----VGGSAGAKK--- 1248
hsE15 SRMSSTKSSGSESDRSERPVLVRQSTFIKEAPSPTLRRKLEESASFESLSPSSRPASPTR 1822
mmE15 SRMSSTKSSGSESDRSERPALVRQSTFIKEAPSPTLRRKLEESASFESLSPSSRPDSPTR 1824
frE15 -----------QKTEPEKPALVRQSTFIKEAPSPTLKRKLEESA----------PAAPSE 1287
hsE15 SQAQTPVLSPSLPDMSLSTHSSVQAGGWRKLPPNLSPTIEYNDGRPAKRHDIARSHSESP 1882
mmE15 SQAQTPVLSPSLPDMSLSTHPSVQAGGWRKLPPNLSPTIEYNDGRPTKRHDIARSHSESP 1884
frE15 --------SPTSPDIFLPSAG----------RRHDVNRSHSESP 1313
hsE15 SRLP------INRSGTWKREHS----------KHSSSLPRVSTWRRTGSSSSILSASSES 1926
mmE15 SRLP------INRAGTWKREHS----------KHSSSLPRVSTWRRTGSSSSILSASSES 1928
frE15 SRPQEATSSRFSRTGTWKRENSSTGAGGGSAGKHSTSLPRVGTWKRTGSSSSVLSASSES 1373
hsE15 SEKAKSEDEKHVNSISGTKQSKENQVSAKGTWRKIKENEFSPTNSTSQTVSSGATNGAES 1986
mmE15 SEKAKSEDERHVSSMPAPRQMKENQVPTKGTWRKIKESDISPTGMASQSASSGAASGAES 1988
frE15 SEKGRSEED--------------GSLRSKGTWRKTK---------SSGGDSSAGRGGS-- 1408

Figure. 5.8. Clustal W aligment of the Apc1 amino acid sequences. Alignment of the human, mouse and *Fugu* Apc1 translated exons. Specific motifs are shown: oligomerisation domain (grey box), armadillo region (yellow box), 15aa repeat (red box), 20aa repeat (green box), basic domain (dark green box), EB1 binding site (pink box), HDLG binding site (blue box) and SAMP repeats (blue characters)..

5.3.5.2. Amino Acid sequence of the *Fugu* Apc2

The *Apc2* gene codes for 14 coding exons, the deduced frApc2 amino acid sequence from exon 8 to 14 extends over 2148aa, although the length of each translated exon is well conserved with the human and mouse exons, exon 14 is 155aa longer than the human and 137aa longer than the mouse exon 14 (Table 5.6).

The first *Fugu* exon codes for 40aa, this is because the sequence that codes for the first seven amino acids is not present in the scaffold, however, the amino acids present in the exon show 90% conservation with the mouse and human exon. The remaining exons are well conserved in the three species. In the Apc2, most of the specific domains are encoded by the exon 14; the overall conservation of the exon is 59.47%.

Exon	Num. of aa conserved	Percentage of conservation
1	36/40	90.0%
2	29/30	96.66%
3	-	-
4	-	-
5	-	-
6	-	-
7	-	-
8	105/133	78.94%
9	30/33	90.90%
10	41/45	91.11%
11	26/26	100%
12	37/39	94.87%
13	71/73	97.26%
14	1070/1799	59.47%

Table 5.6. Amino acid conservation of the Apc2. Alignment of the aa sequence of each exon was done to estimate the number of aa conserved and percentage of conservation among human, mouse and *Fugu*. The *Fugu* aa sequence was used to estimate the level of conservation.

The Apc2 protein is very closely related to its paralogue Apc1, six of the seven main domains of the protein are present in the Apc2 protein with exception of the 15aa repeat (van Es *et al.* 1999), as mentioned earlier, it contains 14 exons, the last exon encodes for most of the specific domains of the protein. We used the alignment produced for each translated exon and the domains characterized in the mouse Apc2

to identify the specific protein domains of the frApc2. Table 5.7 shows the number of amino acids and the degree of conservation (in %) of each domain between the human, mouse and *Fugu*. Interestingly, none of the specific domains of the protein are located in the exons 3 to 7, which are the missing exons in the frApc2.

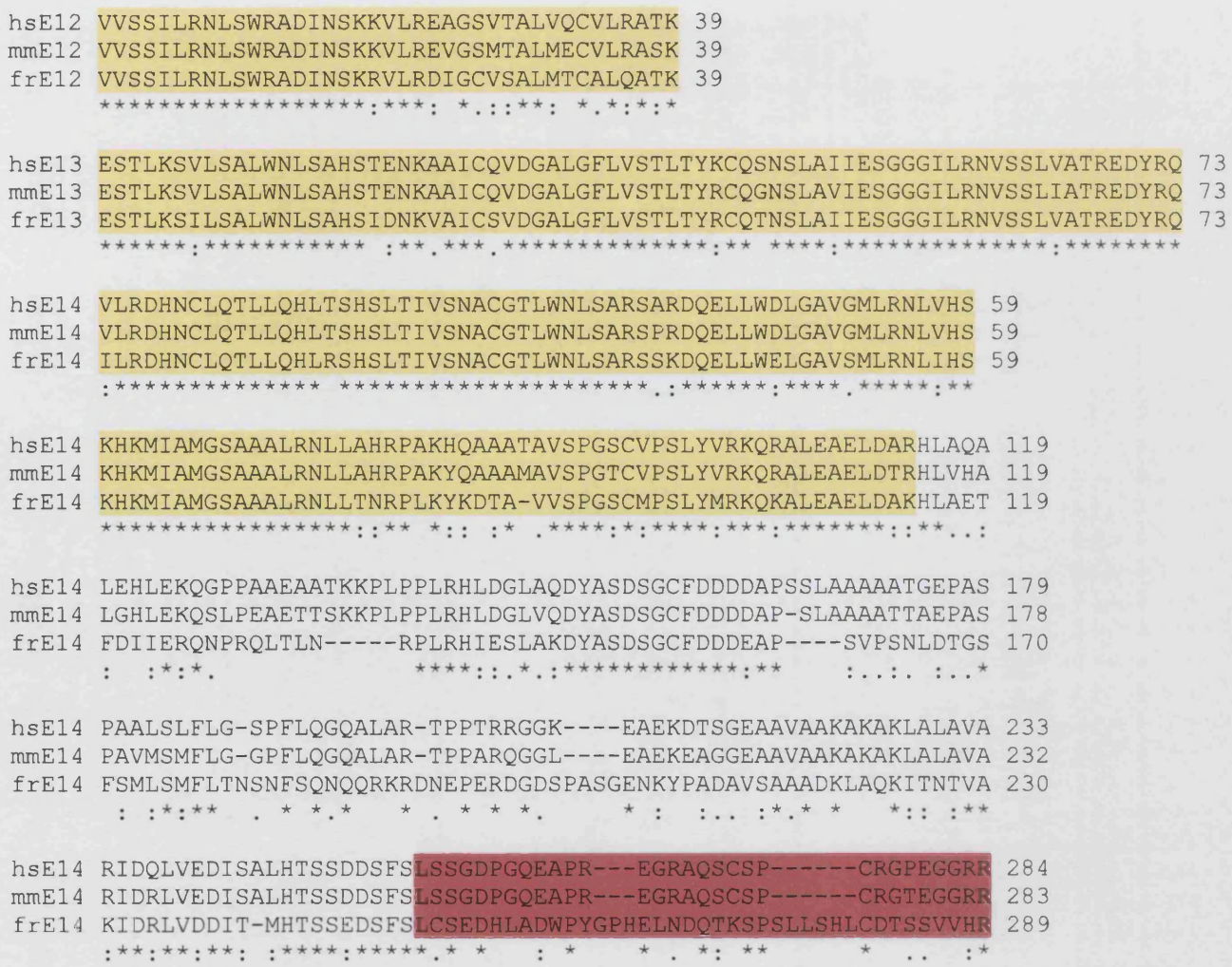
The oligomerisation domain is encoded within the first and second exons of the gene; the exons 9, 10, 11, 12, 13 and 14 encode the seven armadillo repeats and the remaining specific domains are encoded by exon 14. The basic domain, the EB1 binding site and the HDLG binding site show a low level of conservation (50.50, 44.81 and 50% respectively). The low level of conservation shown by the *Fugu* EB1 binding domain is due to the difference in length, it is 21aa longer than the human and mouse domain. The Apc2 20aa repeats also contain the same TPXXFSXXXSL characteristic signature, however only six repeats were identified within this core sequence (Figure 5.9).

Amino acid	Domain	Num. of aa conserved	Percentage of conservation
6- 57	Oligomerisation domain	47/51	92.15%
453- 767	Armadillo region	303/314	96.49%
1020- 1170	15 aa repeat	-	-
1265- 2035	20 aa repeat	13/14	92.85%
		14/14	100%
		8/19	42.10%
		15/16	93.75%
		11/12	91.75%
		11/15	73.33%
2200- 2400	Basic Domain	101/200	50.50%
2559- 2771	EB1 binding site	95/212	44.81%
2772- 2843	HDLG binding site	36/72	50.0%

Table 5.7. Specific domain conservation of the Apc2. Specific domains were delimited using the described domains in the mouse Apc2. The number of aa and percentages are the conservation among human, mouse and *Fugu* domains.

The two *Fugu* SAMP repeats are also located after the fourth and seventh 20aa repeat; the first repeat is 29aa in length, the same length than human and mouse, it is 89.65% conserved (26/29aa) with the mouse and human repeat (Figure 5.9). The second *Fugu* SAMP repeat is three amino acids shorter than the mouse and human, it is 65.38% conserved (17/26aa).

```
hsE1 MASSVAPYEQLVRQVEALKAENSHLRQELRDNSSHLSKLETETSGMK 47
mmE1 MTSSMASYEQLVRQVEALKAENTHLRQELRDNSSHLSKLETETSGMK 47
frE1 -------YDQLAHQVAALRKENCHLRRELEDNSNQLSKLETETFGMK 40
           *:**.:** **: ** ***:**.***.:******** ***

hsE2 EVLKHLQGKLEQEARVLVSSGQTEVLEQLKAL 32
mmE2 EVLKHLQGKLEQEARVLVSSGQTEVLEQLKAL 32
frE2 EVLKQLQSKLEQEAGTLASSGRSDVLHQLK-- 30
     ****:**.****** .*.***:::**.***

hsE8 VEVVFWLLSMLATRDQEDTARTLLAMSSSPESCVAMRRSGCLPLLLQILH----GTEAAA 56
mmE8 VEVVFWLLSMLATRDQEDTARTLLAMSSSPESCVAMRRSGCLPLLLQILH----GTEAGS 56
frE8 VEMVFWLLSILTNRDKDEMSRTLLALSSSPDSCIAMRKSGCVPLLVQILHDGSGGTSSLE 60
     **:******:*:.**::: :*****:****:**:***:***:***:****    **.:

hsE8 GGRAGAPGAPGAKDARMRANAALHNIVFSQPDQGLARKEMRVLHVLEQIRAYCETCWDWLQARDGGPEGGGAGSAP- 132
mmE8 VGRAGIPGAPGAKDARMRANAALHNIVFSQPDQGLARKEMRVLHVLEQIRAYCETCWDWLQARDSGTE-----SAP- 127
frE8 QSASCATVTGCSREAKSRASATLHNIIFCQQDEGQARREKRVLLLLEQIRTHCDTGWDWIENHTVTSTG----TRTN 133
      . :  . :  :::*: **.*:****:*.* *:* **:* *** :*****::*:* ***:: :   .      : .

hsE9 ---IPIEPQICQATCAVMKLSFDEEYRRAMNELGG 32
mmE9 ---VPIEPQICQATCAVMKLSFDEEYRRAMNELGG 32
frE9 DVPEPMDPQICQAVCAIMKLSFEEEYRHALKELG- 33
        *::******.**:*****:****:*::***

hsE10 LQAVAELLQVDYEMHKMTRDPLNLALRRYAGMTLTNLTFGDVANK 45
mmE10 LQAVAELLQVDYEMHKMTRDPLNLALRRYAGMTLTNLTFGDVANK 45
frE10 LQVIAELIHLDQETHGMQNDPINMALRRYAGMAMTNLTYGDVVNK 45
      **.:***:::* * * * .**:*:********::****:***.**

hsE11 ATLCARRGCMEAIVAQLASDSEELHQ 26
mmE11 ATLCARRGCMEAIVAQLGSESEELHQ 26
frE11 ATLCSKRSCLQALVAQLASDSEELHQ 26
      ****::*.*::*:****.*:******
```


hsE12 VVSSILRNLSWRADINSKKVLREAGSVTALVQCVLRATK 39
mmE12 VVSSILRNLSWRADINSKKVLREVGSMTALMECVLRASK 39
frE12 VVSSILRNLSWRADINSKRVLRDIGCVSALMTCALQATK 39
hsE13 ESTLKSVLSALWNLSAHSTENKAAICQVDGALGFLVSTLTYKCQSNSLAIIESGGGILRNVSSLVATREDYRQ 73
mmE13 ESTLKSVLSALWNLSAHSTENKAAICQVDGALGFLVSTLTYRCQGNSLAVIESGGGILRNVSSLIATREDYRQ 73
frE13 ESTLKSILSALWNLSAHSIDNKVAICSVDGALGFLVSTLTYRCQTNSLAIIESGGGILRNVSSLVATREDYRQ 73
hsE14 VLRDHNCLQTLLQHLTSHSLTIVSNACGTLWNLSARSARDQELLWDLGAVGMLRNLVHS 59
mmE14 VLRDHNCLQTLLQHLTSHSLTIVSNACGTLWNLSARSPRDQELLWDLGAVGMLRNLVHS 59
frE14 ILRDHNCLQTLLQHLRSHSLTIVSNACGTLWNLSARSSKDQELLWELGAVSMLRNLIHS 59
hsE14 KHKMIAMGSAAALRNLLAHRPAKHQAAATAVSPGSCVPSLYVRKQRALEAELDARHLAQA 119
mmE14 KHKMIAMGSAAALRNLLAHRPAKYQAAAMAVSPGTCVPSLYVRKQRALEAELDTRHLVHA 119
frE14 KHKMIAMGSAAALRNLLTNRPLKYKDTA-VVSPGSCMPSLYMRKQKALEAELDAKHLAET 119
hsE14 LEHLEKQGPPAAEAATKKPLPPLRHLDGLAQDYASDSGCFDDDDAPSSLAAAAATGEPAS 179
mmE14 LGHLEKQSLPEAETTSKKPLPPLRHLDGLVQDYASDSGCFDDDDAP-SLAAAATTAEPAS 178
frE14 FDIIERQNPRQLTLN-----RPLRHIESLAKDYASDSGCFDDDEAP----SVPSNLDTGS 170
hsE14 PAALSLFLG-SPFLQGQALAR-TPPTRRGGK----EAEKDTSGEAAVAAKAKAKLALAVA 233
mmE14 PAVMSMFLG-GPFLQGQALAR-TPPARQGGL----EAEKEAGGEAAVAAKAKAKLALAVA 232
frE14 FSMLSMFLTNSNFSQNQQRKRDNEPERDGDSPASGENKYPADAVSAAADKLAQKITNTVA 230
hsE14 RIDQLVEDISALHTSSDDSFSLSSGDPGQEAPR---EGRAQSCSP------CRGPEGGRR 284
mmE14 RIDRLVEDISALHTSSDDSFSLSSGDPGQEAPR---EGRAQSCSP------CRGTEGGRR 283
frE14 KIDRLVDDIT-MHTSSEDSFSLCSEDHLADWPYGPHELNDQTKSPSLLSHLCDTSSVVHR 289


hsE14 EAGSRAHPLLRLKAAHASLSNDSLNSGSASDGYCP-REHMLPCPLAALASRR-------- 335
mmE14 EAGSRAHPLLRLKAAHTSLSNDSLNSGSTSDGYCT-REHMTPCPLAALAEHR-------- 334
frE14 DRFSRARALLRLKTAQSSLSTDSLNSGSTSDGYCGSKDQLQPVTRALMMQQRPKQLDLKL 349
hsE14 --------EDPRCG----QPRPSRLDLDLPG-------------CQAEPPAREATSADAR 370
mmE14 --------DDPVRG----QTRPSRLDLDLP--------------SRAELPARDTAATDAR 368
frE14 AHKDYQEADSSLCSNTNRQEIPERDSVDNKHSPVADGGKKQVQPVQTPVSASTKVSSDVS 409
hsE14 VRTIKLSPTYQHVPLLEGASRAGAEPLAG-PGISPGARKQAWLP----ADHLSKVPEKLA 425
mmE14 VRTIKLSPTYQHVPLLDGAAGAGVRPLVG-PGTSPGARKQAWIP----ADSLSKVPEKLV 423
frE14 MASIKLSPSYQQVPSIQSVAKFGVEKASGDVQTAQAMRKQPSVPTAMTAASITKAPTLAS 469
hsE14 AAPLSVAS-KALQKLAAQEGPLSLSRCSSLSSLSSAGRPGPSEGGDLDDSDSSLEGLEEA 484
mmE14 ASPLPIAS-KVLQKLVAQDGPMSLSRCSSLSSLSSTGHAVPSQAENLD-SDSSLEGLEEA 481
frE14 TKSPNIGTMETVQKYSVENTPICFSRCSSLSSLSSGDGALDAQSDNEMESDSSLEIIDVE 529
hsE14 G------------PSEAELDSTWRAPGATSLPVAIPAPRRNRG---------RGLG---VE 521
mmE14 G------------PGEAELGRAWRASGSTSLPVSIPAPQRGRS---------RGLG---VE 518
frE14 DEDLVKKAEEDETLEDLTDSLLLMTDSKSFPSKDTGPVSKTCSSKKEKVFLRAVSPAIVE 589
hsE14 DATPSSSSENYVQETPLVLSRCSSVSSLGSFESPSIASSIPSEPCSGQGSGTISPSELPD 581
mmE14 DATPSSSSENCVQETPLVLSRCSSVSSLGSFESRSIASSIPSDPCSGLGSGTVSPSELPD 578
frE14 DQSPSSSSEHYIHETPLVMSRCSSVSSLGSFESPSIVSSIQSDPCSEMIDGTISPSDLPD 649
hsE14 SPGQTMPPSRSKTPPLAPAPQGPPEATQFSLQWESYVKRFLDIADCRERCRLPSELDAGS 641
mmE14 SPGQTMPPSRSKTP---PAPPGQPETSQFSLQWESYVKRFLDIADCRERCQPPSELDAGS 635
frE14 SPGQTMPPSRSKTPCCLESHVQETQTAGIVSQWEGSLRTFMEIADSKDRLSLPPDLDT-M 708


hsE14 VRFTVEKPDENFSCASSLSALALHEHYVQQDVELRLLPSACPERGGGAGGAGLHFAGHRR 701
mmE14 VRFTVEKPDENFSCASSLSALALHELYVQQDVELRLRPPACPERAVGGGGHRRRDEAASR 695
frE14 IYFTVEKPTENFSCASSLSALPLHEHYIQKDVELKLTPLLQ-----GDKNLLCHDEEGQE 763
hsE14 REEGPAPTGSRPRGAADQELELLRECLGAAVPARLRKVASALVPGR----------RALP 751
mmE14 -LDGPAPAGSRARSATDKELEALRECLGAAMPARLRKVASALVPGR----------RSLP 744
frE14 FDNGERYS----EGNSDDDIEILKECINSAMPSKFRKVRPSLMTQIPPHLVSSQTHKPIH 819
hsE14 VPVYMLVP---------------APAPAQEDDSCTDSAEGTPVNFSSAASLSDETLQGPP 796
mmE14 VPVYMLVP---------------APARG--DDSGTDSAEGTPVNFSSAASLSDETLQGPS 787
frE14 LPVYMMFPNGKTQICPGRKVIVSHKDLKLDDSSLTDSAEGTPGNFSTTTSLSDDTLQYPV 879
hsE14 RDQPGGPAGRQRPTGRPTSARQAMGHRHKAGGAGRSAEQSRGAGKNRAGLELPLGRPPSA 856
mmE14 RDKPAGPGDRQKPTGRAAPARQTRSHRPKAAGAGKSTEHTRGPCRNRAGLELPLSRPQSA 847
frE14 KHRGSKDCLSTSIIKEQELDDEKRIEDLRIFSHFHKLNRTNNPGTQKNRHITPTQRVLMQ 939
hsE14 P---ADKDGSKPGRTRGDGALQS-LCLTTPTEEAVYCFYGNDSDEEPP------------ 900
mmE14 R---SNRDSSCQTRTRGDGALQS-LCLTTPTEEAVYCFY--DSDEEPP------------ 889
frE14 SKEVADRVASQRNRDRSPNQQKNRPCQDLVRNLALPVIMKSDQDAGFQKRNIICKKIAHF 999
hsE14 AAAPTPTHRRTSAIPRAFTRERPQGRKEAPAPSKAAPAAPPP-----ARTQPSLIADETP 955
mmE14 ATAPPP--RRASAIPRALKREKPAGRKET--PSRAAQPATLP-----VRAQPRLIVDETP 940
frE14 LHDNNYVCEDSNANSDAFGQTTRKQTREANAANTGKKENSQGKHRGFQRIKQSLLKDESM 1059
hsE14 PCYSLSSSASSLSEPEPSE---------------PPAVHPRGREPAVTKDPGPGGGRDSS 1000
mmE14 PCYSLTSSASSLSEPEAPE---------------QPANHARG--------PEQGSKQDSS 977
frE14 EGYSLSSSLSSLSDAEFEAGKSKAQQTWYKNRQNKTLNAVQQTKPVSIHSQYEEPSSTSS 1119


hsE14 PSPRAAEELLQRCISSALPRRRPPVSGLRRRKPRATRLDE-RPAEGSRERGEEAAGSDRA 1059
mmE14 PSPRAEEELLQRCISLAMPRRRTQVPGSRRRKPRALRSDI-RPTEITQKCQEEVAGSDPA 1036
frE14 VSMDSEDDLLQKCITSAMPKQRR---KHAARKKKAMNSDKGKKALDAWKLEEELDSDDAD 1176
hsE14 SDLDSVEWRAIQEGANSIVTWLHQAAA-ATREASSESDSILSFVSGLSVGSTLQPPKHRK 1118
mmE14 SDLDSVEWQAIQEGANSIVTWLHQAAAKASLEASSESDSLLSLVSGVSAGSTLQPSKLRK 1096
frE14 SDLNSVEWRAIQEGAN----WLSLGCRPQNLKSLPLKRLNRSFHSCQPLASRLKKGSFAK 1232
hsE14 GRQAEGEMGSARRPEKRGAASVKTSGSPRSPAGPEKPRGTQKTTPGVPAVLRGRTVIYVP 1178
mmE14 GRKPAAEAGGAWRPEKRGTTSTKINGSPRLPNGPEKAKGTQKMMAGESTMLRGRTVIYSA 1156
frE14 ----------TKPVPNFPVVFRGRTVIYTP 1252
hsE14 SPAPRAQPKGTPGPRATPRKVAPPCLAQPAAPAKVPSPGQQRSRSLHRPAKTSELATLSQ 1238
mmE14 GPASRTQSKGISGPCTTPKKTGTSGTTQPETVTKAPSPEQQRSRSLHRPGKISELAALRH 1216
frE14 ------RKETAPSQRPPPTKLTTS-----SDPPKNPNLAQHRSKSLHRLG-HSQDMDLAL 1300
hsE14 PPRSATPPARLAKTPSSSSSQTSPASQPLPRKRPPVT----QAAGALPGPGASPVPKTPA 1294
mmE14 PPRSATPPARLAKTPSSSSSQTSPASQPLPRRSPLAT----PTGGPLPGPGGSLVPKSPA 1272
frE14 PKRSSTPPPRMQKSSSSGSSQTSTPSKQKKTTSLSGTNKNIPKKGVTPTNGPPAAERTAG 1360
hsE14 RTLLAKQH----KTQRSPVRIPFMQRPARRGPPPLARAVP-EPGPRGRAGTEAGPGARGG 1349
mmE14 RALLAKQH----KTQKSPVRIPFMQRPARRVPPPLARPSP-EPGSRGRAGAEGTPGARGS 1327
frE14 AALNSDEKPSTPKTQKSPVRIPFMQNPVK--PRPLSPLVTNQTAGKPANMVHGKVVSPAS 1418
hsE14 RLGLVRVASALSSGSESSDRSGFRRQLTFIKESPG-LRRRRSELSSAESAASAPQGASPR 1408
mmE14 RLGLVRMASARSSGSESSDRSGFRRQLTFIKESPGLLRRRRSELSSADSTASTSQAASPR 1387
frE14 RLELLRMTSAG----RESDRNGFLRQMTFIKESKT-----VQKHDSAKCSMFRSQKRPLH 1469

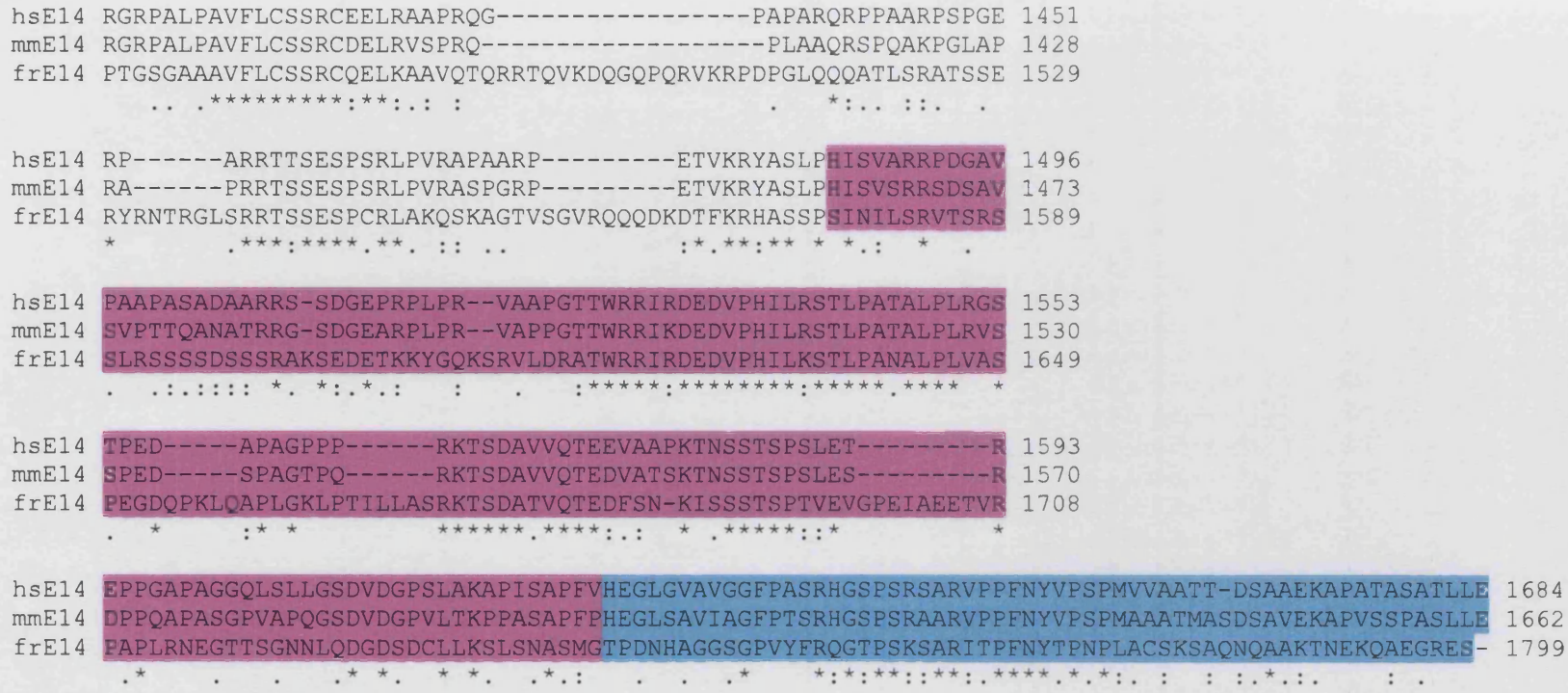


Figure. 5.9. Clustal W aligment of the Apc2 amino acid sequences. Alignment of the human, mouse and *Fugu* Apc2 translated exons. Specific motifs are shown: oligomerisation domain (grey box), armadillo region (yellow box), 20aa repeat (green box), basic domain (dark green box), EB1 binding site (pink box), HDLG binding site (blue box) and SAMP repeats (blue characters).

5.4. Discussion

The tumour suppressor gene *Apc1* codes for a multifunctional protein that performs both signalling and structural functions. One of the main roles of this protein is in the Wnt pathway, where it acts as a component of the β-catenin suppressor complex together with the axin and GSK-3 proteins. *Apc*1 is also involved in cell adhesion and maintaining the internal cell scaffold. The second member of the *Apc* family is *Apc2*, which shares most of the functional domains of the protein and is able to down regulate cellular β-catenin (Nakagawa *et al.* 1998).

Using a comparative genomics approach, the *Fugu Apc1* and *Apc2* genes were identified from the *Fugu* genomic database. The transcriptional organisation of both genes shows no conservation with the mouse and human genomic organisation. However, gene structure, nucleotide and amino acid sequences are well conserved, especially the functional domains of the protein. Moreover, the tissue distribution of the genes seems to be well conserved across species.

Niether of the *Fugu Apc* genes has any synteny with either of the mammalian *Apc* genes containing genomic regions. However, there is also no synteny between the mouse and human *Apc2* containing genomic regions. Most importantly, the intergenic distances in both *Fugu Apc* genes are enormously reduced compared to the mammalian genes; this feature is useful to locate regions involved in the regulation of the genes.

The gene structure of the *Fugu Apc* genes is highly conserved when compared to the vertebrate *Apc* genes, especially for the *Apc1*, where the exon length and intron exon phases are well preserved. For the existing *Apc2* exons that were characterised there is also high degree of homology with the mouse and human.

We were unable to identify the complete *Apc2* sequence using a comparative genomics approach, however, our finding of the first two coding exons would facilitate the completion of the gene; specific PCR primers can be designed for remaining exons and complete the structure and coding sequence of the gene.

Previous studies have identified the expression of *Apc* genes in a wide range of foetal and adult tissues; both genes are highly expressed in whole brain and foetal

brain, and both of them have similar levels of expression in the tissues where they are expressed (van Es *et al.* 1999). Expression analysis by RT-PCR shows that *Fugu Apc1* is expressed in all the adult tissue and embryo stages tested; this expression matches with the pattern of expression reported for mouse *Apc1*. The expression analysis of the *frApc2* reveals expression of the gene during development in the whole embryo, which agrees with the expression reported for the human *APC2* that is expressed in the foetal brain. The expression obtained in the heart also concurs with the expression reported for *hAPC2*. However, van Es et al. (1999), by probing dot blots of normalized poly (A) mRNA, reported a very high expression of the *hAPC2* transcript in the spinal cord and a low level of expression in the small intestine and foetal kidney. In contrast with these results, we did not obtain expression in these *Fugu* adult tissues. Although we found expression in the whole embryo, it could be that the *Apc2* is expressed only in the developing kidney and stops once the organ completes its formation. With respect to the spinal cord and gut tissue, we have not found a possible explanation. However, our results of the kidney and gut expression agree with the results reported in the MGI database for the mouse *Apc2* (see Table 5.3).

Studies performed in the N-terminus of the protein showed that the first 177aa are sufficient and the first 55aa of the protein are necessary for the formation of homodimers (Su *et al.* 1993). This region is encoded by the first and second exons in both the *Apc* genes, which is highly conserved in *Fugu*. The armadillo repeats, which are coded by the exons 10 to beginning of 15 in the *frApc1*, and the exons 9 to beginning of 14 in the *frApc2*, are involved in nuclear transport, cell cycle control and microtubule stability. They have an overall conservation of more than 96% when compared with the human and mouse. The high degree of conservation may indicate that the *Fugu* Apc proteins carry out the same activities of the mouse and human proteins. The armadillo repeat also binds to the Rac-specific guanidine nucleotide exchange factor Asef (Kawasaki *et al.* 2000); this complex regulates the actin cytoskeleton, cell morphology and migration.

The three 15aa repeats and the seven 20aa repeats bind to β-catenin, although only the 20aa repeats seem to be essential for this interaction. The frApc1 15aa repeats have a low degree of conservation compared with mouse and human, whereas this domain is not present in the Apc2 proteins.

The 20aa repeats are well conserved in the frApc proteins, suggesting that the function of the domains in *Fugu* may be in the Wnt pathway, regulating the levels of β-catenin in the cell. The SLSSL motifs in the 20aa repeats are well conserved in the *Fugu* APC genes; these motifs have been involved in the intercellular cell adhesion. The nuclear signal LXXL/I/M/V that are contained in the 3rd, 4th and 7th 20aa repeats in human and mouse are also present and conserved in the frApc1 and the 7th repeat in the frApc2.

The three and two SAMP repeats of the frApc1 and frApc2 respectively are placed in the same positions than in the human and mouse, with around 80% of conservation for most of the repeats. The SAMP domains act in the binding of axin and axin-conductin, a complex necessary for the inhibition of the Wnt pathway.

In human and mouse, the basic domain involved in microtubule association is a very rich region of proline (P), arginine (R) and lysine (K) residues (Polakis 1997). Interestingly, the frApc1 basic domain seems to lack these residues (see Figure 5.8), whereas the frApc2 retains them (see Figure 5.9).

Finally, the EB1 and hDLG domains contained in the C-terminal end of the APC proteins are much better preserved in the frApc1 than in the frApc2. The EB1 site is associated to the binding of microtubules during mitosis, and the hDLG site is involved in cell-cell contact. As in the N-terminal, the conservation of the *Fugu* C-terminal suggests that the frApc proteins perform the same functions in the cell that the mammalian Apc proteins.

In general, the *Fugu* Apc proteins are well conserved between the human and mouse proteins, mainly in the functional domains of the proteins. Although functional assays are necessary to confirm the functions of the *Fugu* Apc, the high conservation of the gene and protein, and the tissue distribution showed by the *frApc* genes strongly suggests that the role of these proteins in *Fugu* is exactly the same as the vertebrate proteins. This high degree of conservation indicates that the regulation of the genes may be conserved across species, which means that the regulatory region and transcription factors involved in this regulation may be preserved. In addition, the analysis of the transcriptional organisation of the *frApc1* already shows that the intergenic distances are much reduced in *Fugu* than in mammals, facilitating the screening for conserved regulatory regions in the non-coding sequences of the gene.

Chapter Six

Identification of the Cis-regulatory elements of *Apc1* in mouse and *Fugu*

Regulation by Cdx transcription factors

6.1 Introduction

The Cdx1 and Cdx2 homeodomain proteins are key players in the regulation of the differentiated intestinal epithelium along the crypt villus axix. The Cdx2 protein activates and regulates the expression of genes involved in cell proliferation and differentiation such as sucrase isomaltase, carbonic anhydrase 1, apolipoprotein B, the pancreatitis associated protein I (PAPI), cyclin D1 among others (Suh *et al.* 1994; Drummond *et al.* 1996; Lee *et al.* 1996). The Adenomatous Polyposis Coli, a multifunctional protein involved in cell adhesion, cell migration and in the canonical Wnt signalling pathway, is also expressed in intestinal epithelium. APC is often mutated in colorectal cancer, where truncated versions of the APC protein are unable to down regulate β-catenin, allowing this protein to enter into the nucleus, form a transcriptional complex with TCF/Lef transcription factors, and as a complex activate a repertoire of genes.

APC is also expressed in a variety of tissues in the central nervous system such as olfactory bulb, hippocampus, brain stem, spinal cord, dorsal root ganglia and cerebellum; and it has also been associated to brain tumours (Senda *et al.* 1998). Many alternative transcripts have been identified for the *Apc*, however, two main transcripts of 10 and 12Kb in length have been identified in brain with approximately similar level of expression in this tissue. The 10Kb transcript is also expressed in lung, liver, skeletal muscle, testis and kidney; the 12Kb transcript is also expressed in these tissues, but in a lower intensity (Wedgwood *et al.* 2000).

Although several studies have addressed the role of *Apc* in the maintenance of the intestinal epithelium and its function in colorectal cancers, little is known about the regulation of the *Apc* gene in the intestinal epithelium.

It is likely that the *Apc* gene is positively regulated by Cdx2 and negatively regulated by Cdx1 in the intestinal epithelium in accordance with the expression pattern and function that each Cdx factor plays in the crypt villus intestinal axis. *Apc* and *Cdx2* exhibit the same expression gradient in the intestinal epithelium (James *et al.* 1994; Jette *et al.* 2004); APC and Cdx2 are able to regulate the expression of the retinol dehydrogenase L (*RDHL*), an enzyme required for the conversion of retinol into retinoic acid; retinoids induce cell differentiation, decrease cell proliferation and increase cell adhesion (Jette *et al.* 2004). Although the mechanism in which *RDHL* is

regulated is unclear, this may not be via Apc regulating *Cdx2*; studies show that familial adenomatous polyps (FAP) tissues express normal levels of Cdx2 (Hinoi *et al.* 2001). Inactivation of *Cdx2* or *APC* results in the development of multiple intestinal polyp lesions (Beck *et al.* 1999) and *Apc*+/- *Cdx2*+/- mice show more colonic polyps than the *Apc*+/- *Cdx2*+/+ mice (Aoki *et al.* 2003).

Combinations of different transcription start sites make the *Apc1* gene a difficult target to study its regulation. A variety of transcripts have been identified in human colon and brain and different spliced forms have been described in mouse, rat and human (Santoro and Groden 1997). Transcripts containing combinations of different untranslated exons have also been identified, such as 0.3, 0.2 and 0.1 exons in combination with exons 1A or 1B (Horii *et al.* 1993). The 0.3/1A transcript was identified to contain a TATA less promoter; reporter assays carrying 100bp upstream of the 0.3 exon showed expression of the reporter in colon carcinoma cells (Thliveris *et al.* 1994). Recently, a new *mApc1* promoter was localized about 40Kb upstream from the ATG; neither TATA nor CAAT boxes were identified in this promoter, but potential Sp1, AP2, AP1, YY1 and C/EBP binding sites were identified, this promoter was found to be active in a mouse embryonic stem cell line (Karagianni *et al.* 2005).

Using sequence analysis, I have identified Cdx binding motifs in the 5' upstream regionof the *mApc* gene. Promoter activity has already been seen for the sequence located upstream of the exon 0.3 in the human and mouse *Apc* gene. The exon 0.3 is the most 5' exon of the *Apc* gene. We asked whether a 607bp region upstream of the mouse *Apc* contains the necessary elements to drive the expression of the gene in intestinal cells. We performed a series of transfection and transactivation studies in CaCo2 and IEC-6 cells in order to investigate the regulation of *Apc*. And we investigated whether Cdx1 or Cdx2 transcription factors activate or repress expression of *Apc* in intestinal cells. This chapter also describes a comparative analysis performed to locate conserved non-coding regions and TFBS with potential value for the regulation of the *Fugu* and mouse *Apc1*.

6.2 Methods

6.2.1. Transfection and transactivation assays

CAT activity of the promoter constructs was measured relative to the promoterless p*CAT* Basic. p*CAT* Control (Promega), containing SV40 promoter and enhancer sequences, was used as a positive control (Appendix section 4A). A 607bp fragment spanning positions –290 to +317 of *mApc* was subcloned, in the correct orientation, upstream of the *CAT* gene to generate p*Apc*SP (Wedgwood et al., 2000). Transactivation assays were performed using expression vectors containing the cDNA of Cdx1 (CMV-Cdx1), Cdx2 (CMV-Cdx2) provided from Dr. V. Subramanian, and GATA4 (pCDNA-GATA4) from Invitrogen (Paisley, UK). PBSKS2+ (BSK) from Promega (Southampton) was used to level the amount of DNA for each sample and as a negative control for transactivation.

The upstream sequence of the *mApc* gene (GeneBank accession number AF209031) was analysed using the Transcription Element Search System (TESS) program and the Searching Transcription Factor Binding Sites (TFSEARCH) program.

Based on the results of the analysis of the *Apc* upstream sequence, transfection and transactivation assays were planned and performed. Further assays were performed, whose relevance was suggested by the results of the former assays.

6.2.2. Cloning of the upstream sequence of the *Fugu Apc1*

PCR primers were designed based on the upstream sequence of the *Fugu Apc1* gene. The oligonucleotides primers synthesized are: 4.0Kb FW frApc1 (5'-ta gga tcc tcc aaa tcg gtg atc tgt gtc cc-3') to produce a 4.0Kb PCR product, 2.0Kb FW frApc1 (5'-ag gga tcc tcc ga agg gca gaa gtc caa atg-3') to produce a 2.0Kb PCR product and 1.0Kb FW frApc1 (5'-gc gga tcc tcc cac cct gct tta ctg att aa-3') to generate a 1.0Kb PCR fragment. A *BamHI* site was added to these three forward primers. To the reverse primer, RV frApc1 (5'-ta ctcgag tcc gtt gga gga tgc agg ttt ga-3') an *XhoI* site was added. The Bac: b227B17x1 from the *Fugu* genomic Bac library (Elgar *et al.* 1999) was used as a template.

PCR was performed in a Perkin Elmer thermo- cycler. The reaction mix was as follows: 5µl buffer (10X), 3µl $MgCl_2$ (50mM), 1µl dNTPs (10mM), primers 1µl each (10µM), 1µl DNA and 0.5µl ROCHE^{TM} Taq polymerase (1.5 U/µl) and 37.5µl

of H_2O. Amplification was achieved with an initial denaturation at 94^0C for 2mins, followed by 35 cycles of 20sec annealing at 60^0C, 2min extending at 72^0C (2min for the 1 and 2Kb fragments and 4mins for the 4Kb fragment) and 1min denaturing at 95^0C, then a final 5min extension at 72^0C. The products of the reaction were checked by running a 2µl aliquot on a 1% agarose 0.5X TBE gel.

6.2.3. Transgene constructs of the *Fugu Apc1* gene

All PCR products were run in a 1X TAE agarose gel and the band excised and electroeluted. This was then dissolved in 85µl of H_2O, 10µl 10X buffer and 5µl restriction enzyme (5U) and incubated at 37^0C for 2hrs. The digested DNA was then phenol /chloroform extracted and precipitated with ethanol. The fragment was then ligated into the T ends of the pGEM-T vector (Promega). Constructs were sequenced using the T7 and SP6 primers.

The pCS2 CMV-GFP (Muller *et al.* 2002) was modified by excising the CMV promoter using the *SalI/HindIII* sites. Sites were blunted by T4 polymerase (ROCHE) and the vector was religated.

Construct 4Kb *frApc1* GFP was generated by subcloning the *BamHI/XhoI* fragment in pGEM-T into the *BamHI/XhoI* site of the pCS2 GFP^{-CMV}, Construct 2Kb *frApc1* GFP was generated by subcloning the *BamHI/XhoI* fragment in pGEM-T into the *BamHII/XhoI* site of the pCS2 GFP^{-CMV} and Construct 1Kb *frApc1* GFP was generated by subcloning the *BamHI/XhoI* fragment in pGEM-T into the *BamHI/XhoI* site of the pCS2 GFP^{-CMV} (Appendix section 4B).

6.3 Results

6.3.1 Characterisation of the *mApc* regulatory region

We found that Cdx binding sites are present upstream of the *mApc* Transcription Start Site (TSS). Six putative Cdx binding sites were found upstream of the TSS, positions –7, -50, -78, -233, -251 and –282bp; one Cdx binding site was found at position +18 and putative GATA binding sites were found at positions –215, -184 and +60bp relative to the TSS (Figure 6.1). Other transcription factor binding sites found in the 600bp *mApc* region include, Oct1, Ap1, Ap2, HNF1 and Nkx2.

Figure. 6.1. Structure of the upstream region of the *mApc* gene analysed using the TESS and TFSEARCH programs. The yellow boxes indicate the Cdx binding sites and the green box indicates the GATA binding sites, the TSS is underlined and part of the exon 0.3 is indicated in a red box.

6.3.2 Endogenous Cdx activates transcription of *mApc* gene

Transfection assays were used to examine the ability of endogenous Cdx factors to activate transcription of an *mApc-CAT* reporter vector containing two single Cdx motifs upstream of the TSS. CaCo2 cells were transfected with *CAT* reporter vectors, and a β-galactosidase control vector. The reporter showed a 6.47 fold-increase in induction when compared to p*CAT*- Basic (Figure 6.2), indicating that endogenous factors like Cdx could be interacting with the *mApc* promoter region.

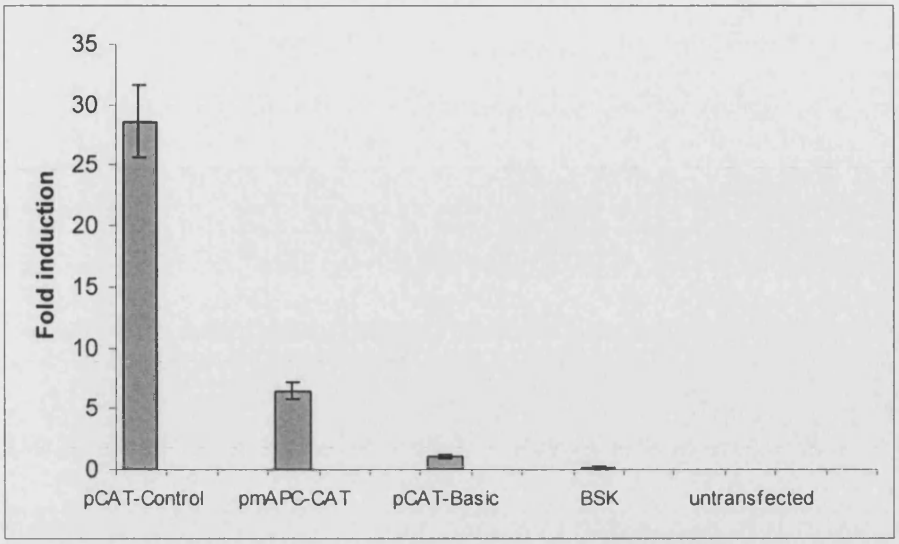


Figure 6.2. Activation of the *mApc* upstream region in CaCo2 cells. CaCo2 cells were transfected with p*CAT*- Basic, p*CAT*- Control, p*mApc*- CAT and pNLs*LacZ* or p27*LacZ* (0.5 μg) as a normalization vector. The –290/+317 *Apc* fragment showed a 6.47 fold induction in comparison with the p*CAT*- Basic vector.

6.3.3 Cdx2 activates the *mApc* regulatory region

Transactivation analysis was performed in IEC-6 cells, which do not express Cdx1 or Cdx2 factors. The IEC-6 cells were transfected with the p*CAT* expression vectors, the CMV-*Cdx2* expression vector and a β-galactosidase control vector. Cdx2 enhanced the reporter transcription by approximately 2.77 fold when compared to the promotorless p*CAT*- Basic (Figure 6.3).

The p*mApc-CAT*/CMV-*Cdx2* normalized values were compared with the p*CAT*-Basic/CMV-*Cdx2* normalized values using a 2-sample t test with unequal variances. The 2 datasets (p*mApc-CAT*/CMV-*Cdx2* and p*CAT*-Basic/CMV-*Cdx2*) followed a normal distribution. The activity showed by the p*mApc-CAT*/CMV-*Cdx2* is significantly higher than the p*CAT*-Basic/CMV-*Cdx2* activity ($t=14.24$, $n=3$, $P=0.005$). This result indicates that Cdx2 is able to activate the *mApc* regulatory region in intestinal cells. This result suggests that other factors are involved in the regulation of *mApc* gene.

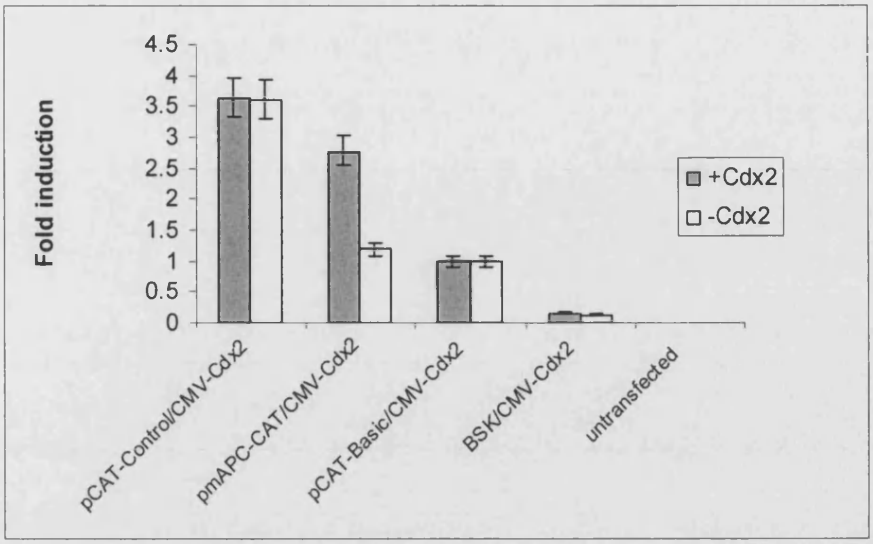


Figure 6.3. Transcriptional activation of *mApc* upstream region by Cdx2. IEC-6 cells were transfected with p*CAT*- Basic, p*CAT*- Control, p*mApc- CAT* (1μg), CMV-*Cdx2* (0.5 μg) and pNLs*LacZ* or p27*LacZ* (0.5 μg). Cdx2 is able to drive the expression of the reporter by 2.77 fold in comparison with p*CAT*- Basic vector. The grey bars indicate the activity of the p*CAT* vectors plus the CMV-*Cdx2* vector; the white bars indicate the activity of the pCAT vectors used without the presence of the CMV- *Cdx2* vector. Each bar represents the average fold of induction of normalized values ± range bars of triplicate points of three independent experiments.

6.3.4 Cdx1 does not activate the *mApc* upstream region

A transactivation analysis was performed in IEC-6 cells. The IEC-6 cells were transfected with the p*CAT* expression vectors, the CMV-*Cdx1* expression vector and a β-galactosidase control vector. Cdx1 did not enhance the reporter transcription when compared to the promotorless p*CAT*- Basic (Figure 6.4). The result of these transactivations suggests that Cdx1 is not acting as an activator of the *mApc* gene in intestinal cells.

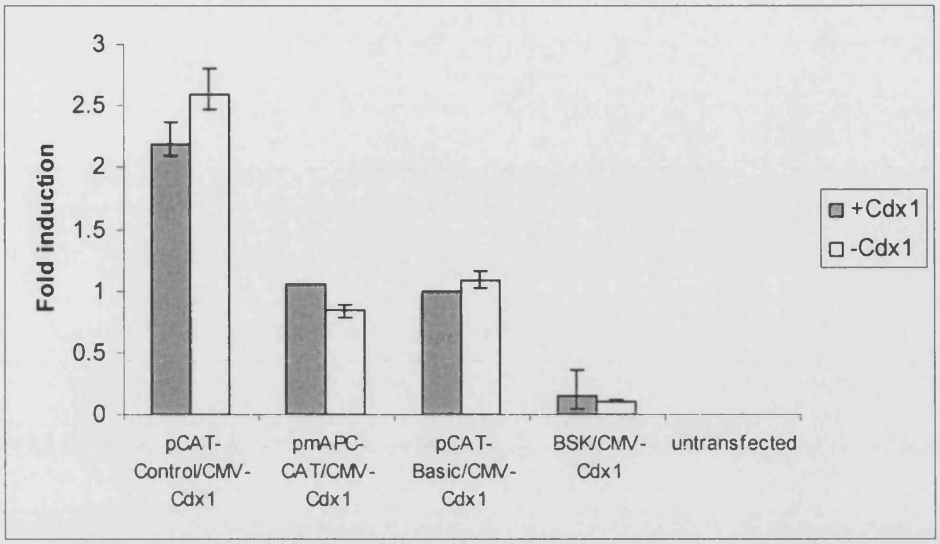


Figure. 6.4. Transcriptional activation of *mApc* upstream region by Cdx1. IEC-6 cells were transfected with p*CAT*- Basic, p*CAT*- Control, p*mAPC- CAT* (1μg), CMV-Cdx1 (0.5 μg) and pNLs*LacZ*, or p27*LacZ* (0.5 μg). Cdx1 vector and p*CAT*- Basic vector show the same expression of the reporter, indicating that Cdx1 is unable to drive the expression of the reporter. The grey bars indicate the activity of the p*CAT* vectors plus the CMV-*Cdx1* vector; the white bars indicate the activity of the p*CAT* vectors used without the presence of the CMV-*Cdx1* vector. Each bar represents the average fold of induction of normalized values ± range bars of triplicate points of three independent experiments.

6.3.5 Cdx1 represses the expression of *mApc* over the activation of Cdx2

A transactivation analysis was done using IEC-6 cells. The IEC-6 cells were transfected with the p*CAT* expression vectors, the CMV-*Cdx2* expression vector, CMV-*Cdx1* expression vector and a β-galactosidase control vector. Cdx1 repressed the activity of the reporter; no induction was obtained when compared to the promotorless p*CAT*- Basic (Figure 6.5). The initial activity showed by Cdx2 (2.77 fold induction) in the second experiment was repressed by the presence of Cdx1, suggesting that Cdx1 may compete for the binding sites present in the upstream sequence of *mApc* and repress the activity of the gene in intestinal cells.

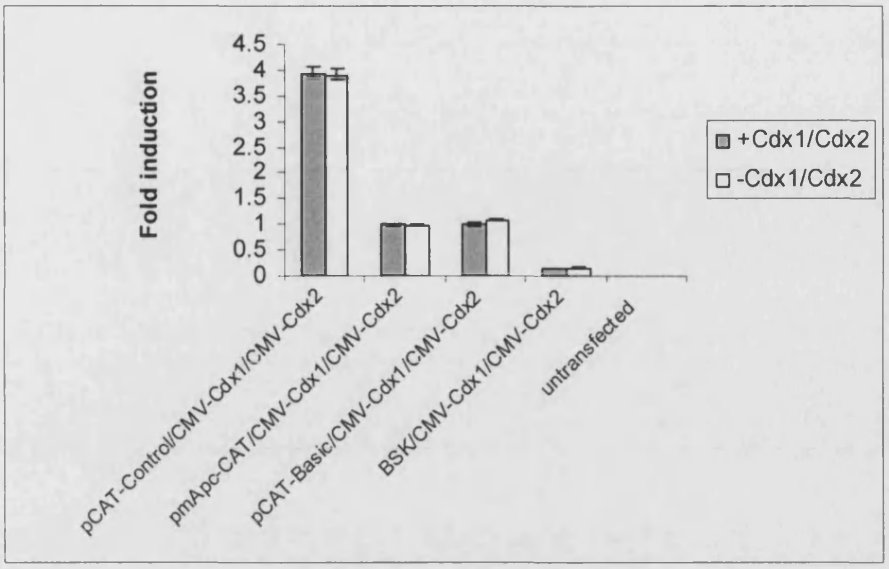


Figure. 6.5. Transcriptional activation of *mApc* upstream region by Cdx1 and Cdx2. IEC-6 cells were transfected with p*CAT*- Basic, p*CAT*- Control, p*mAPC- CAT* (1μg), CMV-*Cdx1* (0.5μg), CMV-*Cdx2* (0.5μg) and pNLs*LacZ* or p27*LacZ* (0.5 μg). Cdx1 repressed the expression of the reporter over the activation of Cdx2, there was no induction compared to the p*CAT*- Basic vector. The grey bars indicate the activity of the p*CAT* vectors plus the CMV-*Cdx1* and CMV-*Cdx2* vectors; the white bars indicate the activity of the p*CAT* vectors used without the presence of the CMV-*Cdx1* and CMV-*Cdx2* vectors. Each bar represents the average fold of induction of normalized values ± range bars of triplicate points of three independent experiments.

6.3.6 GATA4 does not activate the *mApc* upstream region

IEC-6 cells were transfected with the p*CAT* expression vectors, the pCDNA-*GATA4* expression vector and a β-galactosidase control vector. Knowing that GATA4 is an activator transcription factor in the intestinal epithelium, we asked if this transcription factor would be able to enhance the expression of *mApc* in IEC-6 cells to the level of expression shown by the construct in CaCo2 cells. The transactivation assay showed that GATA4 does not enhance the reporter transcription over the promotorless p*CAT*- Basic (Figure 6.6).

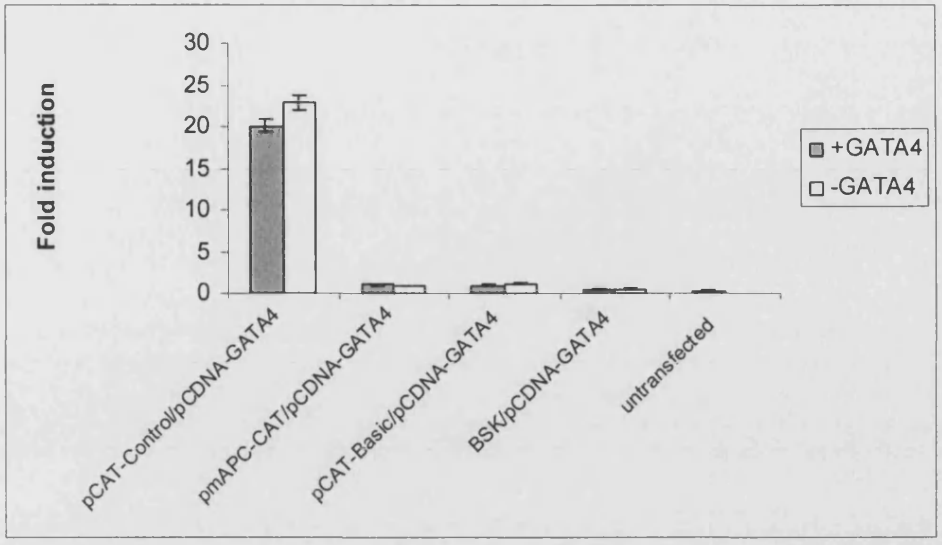


Figure. 6.6. Transcriptional activation of *mApc* upstream region by GATA4. IEC-6 cells were transfected with p*CAT*- Basic, p*CAT*- Control, p*mAPC- CAT* (1μg), pCDNA-*GATA4* (0.5μg), and pNLs*LacZ* or p27*LacZ* (0.5 μg). GATA4 does not activate the expression of the reporter in comparison with p*CAT*- Basic vector. The grey bars indicate the activity of the p*CAT* vectors plus the pCDNA-*GATA4* vector; the white bars indicate the activity of the p*CAT* vectors without the presence of the pCDNA-*GATA4* vector. Each bar represents the average fold of induction of normalized values ± range bars of triplicate points of three independent experiments.

6.3.7 GATA4 may act as a repressor of *mApc*

IEC-6 cells were transfected with the p*CAT* expression vectors, the CMV-*Cdx2* and pCDNA-*GATA4* expression vectors and a β-galactosidase control vector. The reporter activity showed by the p*mApc-CAT* vector was 1.65 fold induction when compared to the promotorless p*CAT*-Basic (Figure 6.7). This result may suggest that GATA4 is able to repress to some extent the expression of *mApc* over the activation of Cdx2 (2.77 fold). A 2-sample t test with unequal variances was used to compare the p*mApc-CAT*/CMV-*Cdx2*/pCDNA-*GATA4* normalized values with the p*CAT*-Basic/CMV-*Cdx2*/pCDNA-*GATA4* normalized values. The 2 datasets followed a normal distribution. The activity showed by the p*mApc-CAT*/CMV-*Cdx2*/pCDNA-*GATA4* is significantly higher than the basal activity (t=3.60, n=3, P=0.037).

However, the comparison between the p*mApc-CAT*/CMV-*Cdx2*/pCDNA-*GATA4* (1.65±0.27 fold) and p*mApc-CAT* (1.45±0.05 fold) do not show a considerable difference in activity. This result suggests that even though the activity of *mApc* was decreased, this may not be due to the presence of GATA4 acting as a transcriptional repressor.

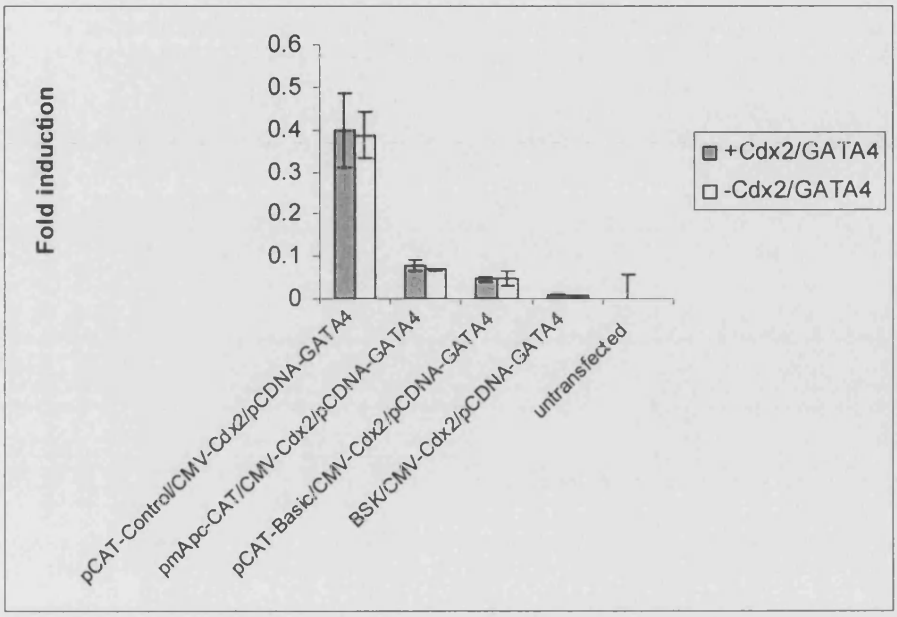


Figure. 6.7. Transcriptional activation of *mApc* upstream region by Cdx2 and GATA4. IEC-6 cells were transfected with p*CAT*- Basic, p*CAT*- Control, p*mApc- CAT* (1μg), CMV-*Cdx2* (0.5μg), pCDNA-*GATA4* (0.5μg), and pNLs*LacZ* or p27*LacZ* (0.5 μg). GATA4 seems to express the reporter expression of *mApc* (1.65±0.27 fold, P=0.037) in comparison with p*CAT*- Basic vector. However, the difference in activity between the *mApc* plus Cdx2/GATA4 and the *mApc* without Cdx2/GATA4 (1.45±0.05 fold) is not significant. The grey bars indicate the activity of the p*CAT* vector plus the CMV-*Cdx2* and pCDNA-*GATA4* vectors; the white bars indicate the activity of the p*CAT* vectors only. Each bar represents the average fold of induction of normalized values ± range bars of triplicate points of three independent experiments.

6.3.8 Cdx1 and GATA4 do not activate the *mApc* upstream region

IEC-6 cells were transfected with the p*CAT* expression vectors, the CMV-*Cdx1* and pCDNA-*GATA4* expression vectors and a β-galactosidase control vector. Cdx1 and GATA4 did not enhance the reporter transcription when compared to the promotorless p*CAT*- Basic (Figure 6.8). The result of this transactivation corroborates that Cdx1 does not act as an activator of the *mApc* gene and GATA4 may not be involved in the regulation of *mApc* when tested in intestinal cells.

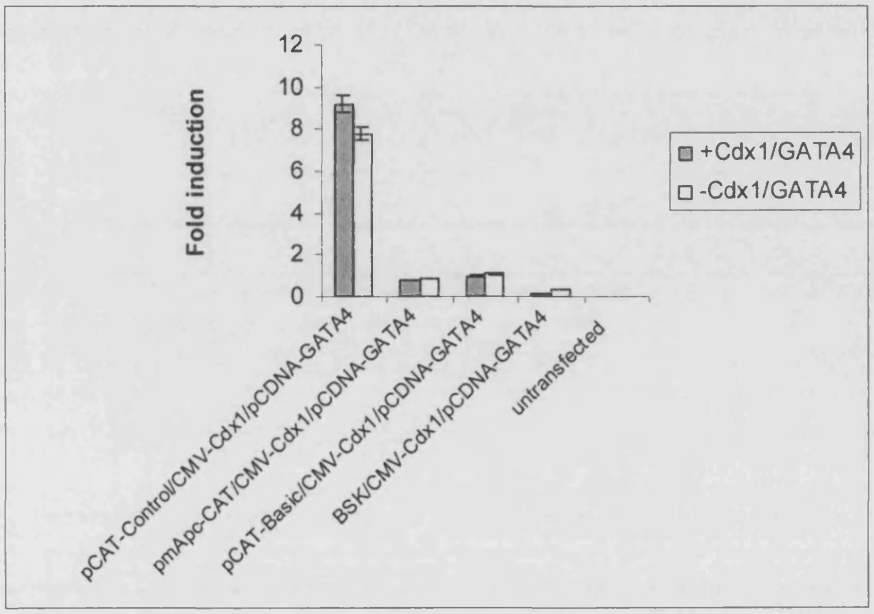


Figure. 6.8. Transcriptional activation of *mApc* upstream region by Cdx1 and GATA4. IEC-6 cells were transfected with p*CAT*- Basic, p*CAT*- Control, p*mApc- CAT* (1μg), CMV-*Cdx1* (0.5 μg), pCDNA-*GATA4* (0.5 μg) and pNLs*LacZ*, or p27*LacZ*. Cdx1 and GATA4 were unable to drive the expression of the reporter in comparison with p*CAT*- Basic vector. The grey bars indicate the activity of the p*CAT* vectors plus the CMV-*Cdx1* and pCDNA-*GATA4* vectors; the white bars indicate the activity of the p*CAT* vectors used without the presence of the CMV-*Cdx1* and pCDNA-*GATA4* vectors. Each bar represents the average fold of induction of normalized values ± range bars of triplicate points of three independent experiments.

6.4. Characterization of the frApc1 upstream region

6.4.1 Analysis of the *frApc1* non-coding region

To identify the conserved non-coding regions in the *Fugu Apc1* upstream region, a comparative analysis was done using the human, mouse and *Fugu Apc1* upstream sequences. The human and mouse *Apc1* upstream sequences were extracted from the Ensembl database; to locate and prepare an annotation file, each sequence was blast against the 0.3 *mApc1* sequence (NCBI accession number: AF209032), producing a 73.25% and 98.42% sequence identity respectively. The *Fugu Apc1* upstream sequence was extracted from the Fugu Databse (M000281, see chapter 5).

A multiple alignment was done using the Mlagan program; the mouse upstream sequence was used as base sequence. The mouse and human 0.3 exon sequences aligned in the predicted position, the *Fugu* upstream sequence failed to produced any significative alignment with the mouse 0.3 exon (Figure 6.9). A multiple alignment using the *Fugu* upstream sequence as a base sequence also failed to predict any conserved *Fugu* 0.3 exon.

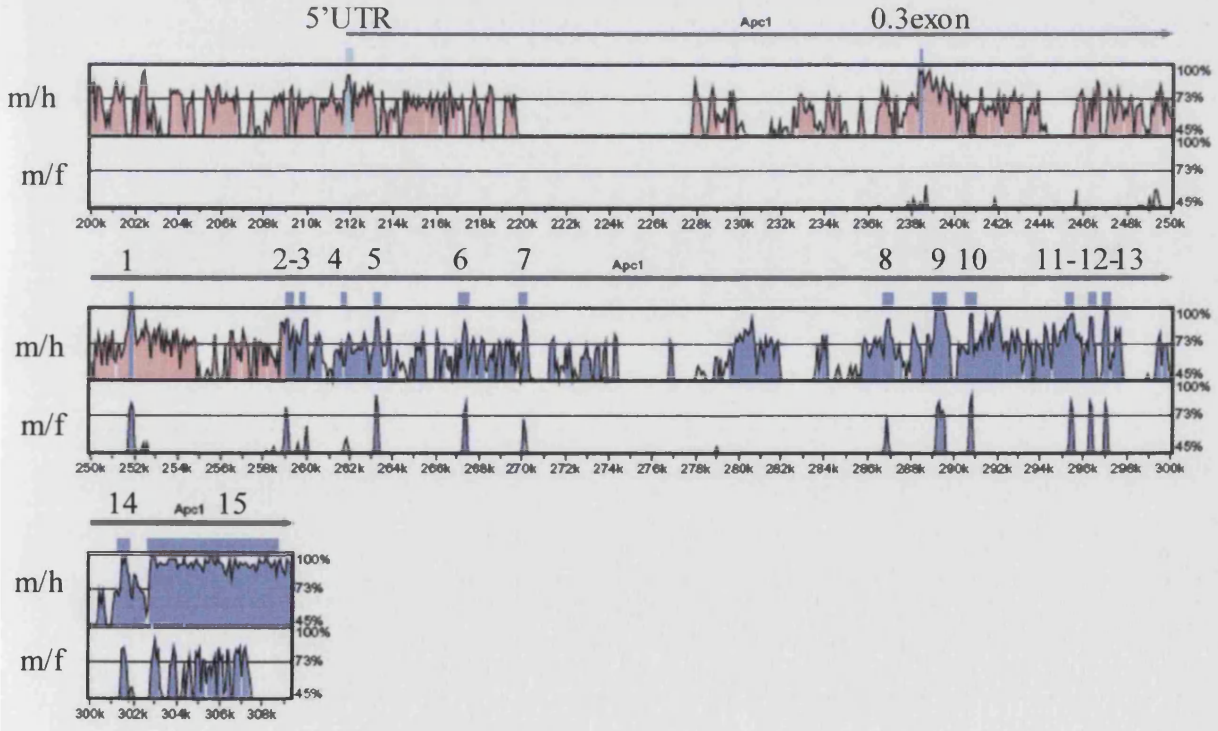


Figure. 6.9. Vista plot representing the *Fugu*, mouse and human upstream and coding sequence alignment. The *mApc1* sequence was used as base sequence for the

alignment. The mouse and human (m/h) contain conserved non-coding region in the upstream and intronic regions of the gene. The mouse and *Fugu* alignment (m/f) failed to produce any conserved non-coding region. The 5'UTR is indicated in light blue, the 0.3 exon and each coding exon are in dark blue.

6.4.2. TFBS in the *frApc1* upstream region

A 4.0Kb *frApc1* upstream region was isolated to investigate putative TFBSs present in the sequence. Based on the literature, potential transcription factors or TF complexes were located according to their position in the GFP constructs (see section 6.2.3.). The predicted transcription factors are summarized in table 6.1.

TFBS in the *frApc1* upstream region			
4.0Kb *frApc1 GFP*		2.0Kb *frApc1 GFP*	1.0Kb *frApc1 GFP*
GATA1/CDX	CEBPβ/CDX/CEBPα	GATA2/GATA1	CDX/CDX/AP1
GATA1/CDX/AP4	CDX/CDX/CEBPβ	SP1/GC box	AP4/CDX/GATA1
GATA2/GATA1/AP4	TATA/CDX/CEBPα	SP1/GATA2/GATA1	GATA1/Sox5/CDX
cEts/cEts/CDX	TATA/GATA/OCT1	AP4/CCAAT/GATA1/AP4	AP4/CDX
Ap1/Oct1/Ap1/CDX/ Ap1	CDX/GATA1/CDX	SP1/GCbox/SP1/AP4	CDX/Sox5/Ets/CDX/ OCT1
CDX/TATA	GATA2/CEBPβ/CDX	GATA1/CDX/GATA2	GATA1/CEBP
AP1/CDX/HNF	GATA2/GATA3	CEBPα/CEBPβ/ GATA2	OCT1/S8/CDX/OCT1
Sox5/Sox5	OCT1/Ets1/Ets1	Sox5	OCT1/CDX
C/EBP	GATA1/CDX/GATA2	CDX/GATA2	CDX/AP4
GATA1/GATA2/AP2/ GATA3	AP4/CDX/GATA1/ GATA2/GATA3/ CEBPβ/CDX	Sox5/SP1	CDX/CDX/OCT1
GATA1/AP4/GATA2/ AP4/GATA1/AP4/ GATA3	OCT1/Sox5/AP4	SP1/GC box	CDX/CDX/TATA
STAT/GATA1	CDX/CDX	SP1/GATA1	Sox5
AP4/GATA2/GATA1		SP1/GATA2/Gcbox/GATA3/CEBP/SP1	CDX/HNF/GATA1/ GATA2
CDX/CDX		AP1/CDX	Ets/SP1/CDX
CDX/OCT1/OCT1		YY1	TATA
CDX/CDX		Ets/GATA1/GATA2/ YY1	
CDX/CDX/CEBP/ CEPBα		GATA1/CDX	
CDX/CDX		CDX/OCT1	
CEBPβ/CDX		AP4/GATA1/AP4	

Table 6.1. The *frApc1* upstream sequence was examined for TFBS; the significant transcription factors were placed in one of the GFP expression constructs (Szpirer *et al.* 1992; Tronche and Yaniv 1992; Kelly and Moon 1995; Zhang *et al.* 1996; Karg H *et al.* 1997; Lopez-Rodriguez *et al.* 1997; Larsson *et al.* 1998; Oates *et al.* 1999; Allan *et al.* 2001; Boudreau *et al.* 2002; Consales and Arnone 2002; Flock and Drucker 2002; Jacobsen *et al.* 2002; van Heel *et al.* 2002; Beland *et al.* 2004; Afouda *et al.* 2005; Jacobsen *et al.* 2005; Seligson *et al.* 2005).

Several TFBS were found in the *frApc1* 5' flanking region. Most of the factors contain developmental regulatory functions in intestinal cells. The distal region (-4 to –2Kb) harbors CDX/CDX, GATA/CDX, Oct/CDX/Ap1 and CEBP/CDX complexes. The region contained between –2 to –1Kb from the TSS, contains almost the same repertoire of binding sites, some GC boxes and YY1 binding sites are present in this element. The TFBS predicted in the proximal upstream region (-1000 to +1bp) are mainly HNF, GATA, CDX, TATA binding sites. Although most of the factors have many binding sites in the full length of the region, the possible combination of factors is different, suggesting that *Apc1* may be regulated by a specific group of transcription factors.

6.4.3. Analysis of the *Fugu Apc1* promoter

The possible absence of the 0.3 exon in the *frApc1* gene may suggest that the promoter region is located nearby the translation start site of the gene. A search for promoter-proximal elements was carried out in 1Kb upstream of genomic DNA (Figure 6.10). The survey of the proximal region revealed two consensus motifs for Sp1 (a GC binding protein) placed at –161 and –789bp form the TSS. Two binding sites for Ap1 positioned at –258 and –487bp from the TSS. One consensus GATA site is predicted at position –280bp and three TATA boxes were located at –304, -598 and –942bp from the TSS. These predicted binding sites suggest that the *frApc1* promoter may be located around this proximal 5' region.

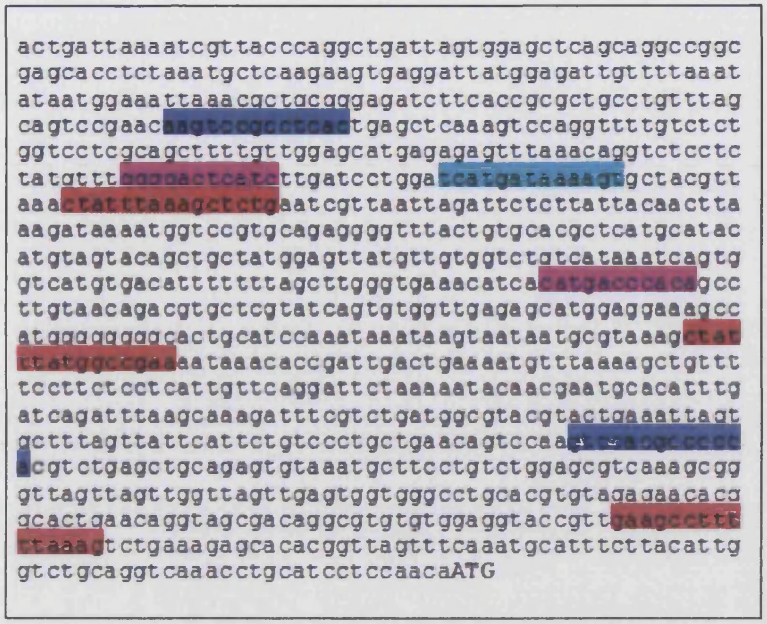


Figure. 6.10. Sequence and putative TFBS in the proximal 5'upstream region of *frApc1*. The translation start site is indicated (ATG). The potential TFBS are color boxed: Sp1 sites (dark blue), AP1 sites (pink), GATA sites (light blue), TATA box (red).

6.4.4. Analysis of the *frApc1* cis-regulatory elements in the developing zebrafish embryo

To investigate if the upstream region of the *Fugu Cdx1* gene is sufficient to control the expression of the gene during development and in intestinal tissue, three GFP expression constructs containing different *frCdx1* non- coding sequence lengths were constructed. Expression constructs were injected into the 1-2 cell-stage embryo. Embryos were analysed at 5 hours post- injection (hpi) for GFP expression.

The *Fugu Apc1* expression constructs were injected at three different concentrations, 15, 25 and 50ng/µl. The three concentrations used produced a high level of mortality in the injected embryos; no survivors were obtained after the gastrulation stage.

6.5 Discussion

We investigated the regulation of *mApc* using a 607bp upstream region of the gene. Transfection assays in CaCo2 cells showed a 6.47 fold induction. Transactivation assays using *Cdx1* or *Cdx2* showed that these factors might regulate

mApc in an antagonistic way. Further analysis of *mApc1* upstream region, showed the presence of GATA factor binding sites. However transactivation assays using GATA4 and combinations of GATA4 and Cdxs factors failed to reveal any involvement of GATA4 in the regulation of the *mApc*.

We started to analyse the *mApc* promoter with transfection analyses using CaCo2 and IEC-6 intestinal cell lines. Initial transfection studies showed that there is activation of the *mApc* promoter in CaCo2 cells (6.47 fold induction), indicating that endogenous Cdx may be involved in the regulation of the gene. Previous studies have used the same *mApc* region in L-cells; the reporter showed an expression of 8.6 fold when compare to the basal activity (Wedgwood *et al.* 2000).

To investigate if the Cdx factors were involved in the regulation of *mApc*, transactivation assays were carried out using *Cdx1*, *Cdx2* cDNA or a combination of both; Cdx2 induced the expression of the reporter to 2.77 fold induction when tested in IEC-6 cells. Due to the expression obtained in CaCo2 cells, this result suggests that Cdx2 may not act alone in the control of *mApc* expression. The activity shown by the reporter when transfected with Cdx1 was null. Moreover, when both Cdx factors were transfected, the induction previously shown by Cdx2 was abolished, indicating that the Cdx factors may be acting in an antagonistic way in the regulation of *mApc*, and that the repressor role of Cdx1 may be stronger than the activator role shown by Cdx2.

The activation of *mApc* shown by endogenous Cdx and the activation obtained in the transactivation assay using Cdx2 suggest that Cdx factors do not act alone in the activation of the *mApc* gene, other factors may be also involved in this regulation.

Studies have shown that the Cdx2 protein can bind consensus AT motifs in either proximal or distal regions of the genes (Suh *et al.* 1994; Troelsen *et al.* 1997; Colnot *et al.* 1998). The Cdx1 factor, also binds to the same consensus site. Due to the similar binding site shared by Cdx1 and Cdx2, the binding mechanism of these proteins might be restricted by other mechanisms than just the binding site itself (Moucadel *et al.* 2001); different factors may be implicated in the interaction with Cdx to discriminate their specific targets (Moucadel *et al.* 2002).

We continued analysing the promoter region of *mApc*, looking for transcription factors that are also specifically expressed in the intestine and have binding sites in the upstream region of *mApc* and may interact with the Cdx factors to regulate

expression. Three putative GATA binding sites were found at position –215, -184 and +60bp relative to the TSS of the gene; the GATA binding site located at position –215bp was found to be 9bp away from one Cdx binding sites, which could set a good spatial scenario for these two transcription to interact if they are involved in the regulation of *mApcI.*

GATA factors have been shown to be expressed during development, cell fate specification, differentiation and proliferation. In contrast to other transcription factor families like the homeobox or winged helix, which have several members, the GATA family only has 6 members in vertebrates (Patient and McGhee 2002). During development, GATA4 is expressed in the primitive and definitive endoderm, and mice lacking GATA4 expression do not have a proper differentiation of the gastric epithelium (Jacobsen *et al.* 2002).

A previous study showed that HNF-1α, GATA4 and Cdx2 are essential for the spatial and temporal expression the Sucrose Isomaltase gene (SI). The HNF-1α protein is present in enterocytes in the villus, with a very reduced expression in the crypt; the GATA4 protein is also expressed in the villus, and no expression is detected in colon. HNF-1α, GATA4 and Cdx2 factors interact with each other to form a complex and regulate the expression of the SI gene (Boudreau *et al.* 2002).

We then asked if GATA4 was involved in the regulation of *mApc*; the transactivation assay showed no induction of the reporter, suggesting that GATA4 may not be involved in the regulation as an activator of the gene. However, when Cdx2 and GATA4 were co-transfected, the 2.77 fold induction showed by Cdx2 was reduced to 1.65 fold induction; at this point GATA4 may be acting as a repressor rather than an activator. However, the values between the *mApc* promoter transfected with GATA4 and Cdx2 (1.65 $\pm$0.013 fold induction) and the Apc alone (1.45$\pm$0.024 fold induction) are not markedly different. Therefore, it is difficult to assign to GATA4 a repressor role in the regulation of *mApc*. Co-transfection of Cdx1 and GATA4 corroborated that these proteins do not activate the expression of the reporter in IEC-6 cells.

Reports have described that GATA proteins can activate transcription of genes via protein complex formation, without a direct contact with the DNA (Merika and Orkin 1995). The GATA5 and HNF-1α interact to activate the lactase promoter, in this case, GATA5 does not make a physical DNA contact (Krasinski *et al.* 2001). It may

be possible that other GATA factors are involved in the regulation of the *mApc* gene without a direct protein DNA interaction.

The results from this study confirm that the immediate upstream region of the *mApc* exon 0.3 is able to induce the expression of the reporter gene, not only in L-cells (as shown by Wedgwood et al., 2000) but also in colon intestinal cells. However, the transactivation assays using Cdx1, Cdx2 or GATA4, or the combination of them were unable to confirm in a consistent way that these factors really participate in the *mApc* regulation. Further studies, for example mutation of the Cdx binding sites or electromobility shift assays, would confirm the binding of these factors to the upstream region of the *mApc*.

The initial activity obtained in CaCo2 cells suggests that the number of transcription factors in the cell may be limited, or that Cdx1 and Cdx2 may act in an antagonistic way. There is also the possibility that other factors apart from the Cdxs and GATA4 are involved in the regulation of *mApc*, such as HNF factors or other GATA factors. Another possibility is that the 600bp fragment might not contain the regulatory elements involved in the expression of this gene in the intestinal epithelium.

In our attempt to characterize the *mApc1* regulatory regions, comparative analyses of the non-coding regions were performed using the *Fugu rubripes Apc1* gene. Comparison of the 5' flanking region of the gene showed low level of conservation between mouse and *Fugu*. Further comparisons revealed that the mouse and *Fugu* upstream sequences shared a common repertoire of TFBS, suggesting a conserved regulatory mechanism between species. These TFBS can be categorized in three main groups: 1) transcription factors expressed in intestinal cells, e.g. Cdx, GATA and HNF; 2) transcription factors involved in developmental processes e.g. Cdx, S8, YY1, Sox5, Oct1 and 3) activator and stimulator proteins e.g. Sp1, Ap1 and CEBP.

The analysis of the *frApc1* proximal upstream region showed the presence of binding sites characteristic of promoter sequences, indicating that *frApc1* promoter may be contained in the first Kb upstream of the translation start site.

Unfortunately, expression assays in zebrafish embryo failed to confirm the relevance of these potential trascription factors. There are two possible explanations for the embryo death produced by the *frApc1* reporter constructs. 1) The concentration used could have been so high that it caused toxicity to the cells at early stages of

development. However, the same range of concentrations and the same expression vector were used in the *frCdx1* expression constructs (see chapter 4) and no toxicity was observed. 2) It could be possible that the *frApc1* reporter constructs are highly expressed in early developmental stages (earlier than 80% epiboly) and caused toxicity to the cells.

In situ experiments indicate that the *ZfApc* is expressed by the 20 somites stage (30-36 hrs) in the zebrafish embryo. Expression is detected in telencephalon, diencephalon, hindbrain, cranial ganglia, tegmentum ventricular zone and spinal cord (Sprague *et al.* 2001). In addition, the RT-PCR data (see chapter 5) shows that *frApc1* is expressed at 53.3 and 111.5hpf in the *Fugu* embryo. This indicates that the *Apc1* gene is expressed during development in both zebrafish and *Fugu*.

There are still some questions to be addressed to fully characterize the upstream region of the *frApc1*; 1) to confirm the presence of the 0.3 exon or any other possible translation start site and 2) obtain the 5'UTR of the gene. Based on the comparative analysis performed here, the presence of these elements (if any) may diverge completely from those in mouse and human to those in *Fugu*.

Chapter Seven

General Conclusions

Conclusion

The development of the organism is controlled by specific and well-structured mechanisms; the expression and regulation of genes involved in this process are tightly regulated from the simplest tissues to the most specialized cells.

Several reports have described the role of transcription factors during development, and their role in specifying the formation of tissues and body structures. In the intestinal epithelium, the Cdx proteins are involved in maintaining the architecture and structure of the CV axis. The *Cdx* genes also intervene during development in specifying the AP axis. Despite of our understanding in the formation and architecture of the intestine, the exact mechanisms in how these processes happen are far from understood.

In this work, we aimed to identify and characterize the regulatory elements involved in the regulation of *Cdx1* and *Apc1*. We made use of a variety of tools from comparative genomics and the use of the *Fugu* genome as a comparative model, to the application of reporter systems and functional gene expression in the developing zebrafish.

Using comparative analyses, we have identified, for the first time, the *Fugu Cdx* family (*frCdx1*, *frCdx2* and *frCdx4*) and the *Fugu Apc1* and *Apc2* genes. The *Fugu Apc* and *Cdx* genes were almost completely characterized. The genomic, gene and amino acid structures of the *Fugu* showed high level of conservation when compared to the human and mouse orthologs. Most of the main and functional domains in the mammalian Apc and Cdx proteins are also contained in the *Fugu* proteins, indicating a conserved functional role across species.

Expression studies of the *frCdx1* and *frCdx2* genes showed that the expression of these genes is restricted to the intestine, as the mouse and human *Cdx* genes are. The *Fugu Apc* genes were expressed during development and in adulthood; some of this specific expression agrees with the expression reported for the mouse *Apc* genes. Specific domains in the *Apc* proteins are present in the frApc proteins, suggesting that their role and function was determined early in the evolution of vertebrates.

More detailed studies were performed with the *Fugu Cdx1* and *Apc1*. Analyses of the 5' flanking region of these two genes, revealed potential binding sites for transcription factors, which may be involved in their regulation. Specially in the *Fugu Apc1* where the reduced and diverged 5' region will provided important information of the structure and regulation of the gene.

The putative transcription factors are involved during development in vertebrae formation, and in intestinal proliferation and differentiation. Strangely, the *frCdx1* and *frApc1* did not show conserved non-coding regions when compared with the human and mouse homologs.

A complete study of the proximal *Fugu Cdx1* upstream region shows a well-conserved set of TFBS with the mouse and human regions. Interestingly, reporter assays performed in intestinal cells showed that the *Fugu* and mouse *Cdx1* upstream regions are able to drive, and in some cases, to enhance the expression of the reporter in the cells. We conclude that the 1Kb upstream region of the mouse and *Fugu Cdx1* contains the necessary elements to drive the expression of the gene; these data agree with reports for the regulation of the mouse *Cdx1*.

Functional characterization of the *frCdx1* cis-regulatory elements in the zebrafish embryo showed that the upstream region of the *frCdx1* also contains elements to drive the expression of the transgene during zebrafish development. Expression of the transgene was detected early during development at the gastrulation stage; later, the expression was restricted to the posterior part of the embryo, in the tail bud.

We investigated if the regulation of the *mApc1* gene is controlled by the Cdx factors in intestinal epithelium. Transfection and transactivation assays confirmed that Cdx1 and Cdx2 regulate to some extent the *mApc1* transcription. Comparative analyses using the *frApc1* revealed that the human and mouse 5' flanking regions have diverged from the *Fugu* upstream region during evolution.

Further analyses will confirm if the potential transcription factor binding sites predicted in the *Fugu* and mouse *Cdx1* and *Apc1* upstream regions are important for the regulation of the genes. More functional assays will prove the specific non-coding regions and the transcription factors involved in the spatial expression of these genes.

The use of comparative genomics comes from the idea that sequences that regulate gene expression are often conserved between species. Because most of the *Fugu* genome has been sequenced and almost completely characterized, and a *Fugu* genome database has been released and is available for use, the *Fugu* genome is a useful model to look for conserved non-coding sequences. Despite this, we were unable to identify conserved non-coding sequences between the mouse and *Fugu* *Apc1* and *Cdx1* genes.

The use of distant orthologous sequences can bring some disadvantages. Studies have shown that the chicken orthologous sequences contain conserved non-coding regions with mouse and human; however, this represents only a subset of elements of the complete set of conserved sequences. Several human-mouse conserved non-coding elements (which are not conserved in chicken) have been confirmed to regulate gene expression. The use of comparative multispecies studies can, sometimes, fail to spot specific mammalian regulatory sequences (Pennacchio and Rubin 2001).

There are different model systems to look for gene functions and functionality of conserved non-coding elements; as mentioned earlier, the system has to satisfy two conditions. First, the regulatory elements to be tested have to be easy to screen. This is sometimes hard to satisfy; some genes with complex patterns of expression contain their regulatory elements dispersed among introns, and sometimes in distant flanking regions. Second, the spatial and temporal expression of the genes or regions tested has to be easy to visualize. This condition can also be difficult to meet, especially for mammalian genes that are expressed in multiple tissues at different developmental stages; cell lines cannot duplicate those patterns in expression, and mammalian models are costly and slow.

We found that the zebrafish can satisfy these two conditions and is a useful model to look for non-coding sequences implicated in development. The fertilization of the zebrafish eggs takes place externally and the embryo development occurs very quickly. Most of the organs' structures and functions are developed by 48hpf. The majority of the developmental processes, signalling pathways and genes are conserved in the teleost through the amniotes. The use of the green fluorescent protein as a transgene in the zebrafish embryo has been a useful tool for developmental and

genetic studies; the product of GFP is very easy to track in the clear zebrafish embryo, allowing us to follow the spatial and temporal expression of genes.

In this work, we have identified and characterized the *frCdx* and *frApc* genes. We analysed and functionally characterized the upstream sequence of the *frCdx1* and *frApc1* genes. The data presented here will contribute to complete our understanding in the characterization and function of non-coding regions (in some cases conserved across species, in some cases not), which contain a potential value in the regulation of genes.

This study also tried to complete the knowledge that we have of the regulation of specific intestinal genes, one a transcription factor (Cdx1), the other a signaling protein (Apc1). Both perform different functions in the cell, but contribute to maintain the architecture and structure in the intestinal epithelium.

The regulation of genes is not a linear interaction between one transcription factor and the DNA, or between two proteins. It is a complex mechanism where molecules, signals and proteins are in a cross-talk communication.

A complete understanding of how architecture and structure are maintained in the regulatory systems is only possible through the understanding of the regulation of all its components and the interactions amongst them. A gene regulatory network representing the transcription factors and signaling molecules involved in the formation and specification of the endoderm and mesoderm in sea urchin has been created using a variety of techniques (e.g. bioinformatics, genomic data, cis-regulatory analyses and molecular embryology). This network is only one component of a more complex system; each component in the system performs different (and related) biochemical and cellular processes in the cells, but these tasks are given from a genomic regulatory control system (Davidson *et al.* 2003; Davidson 2004).

Finally, we tried to study and increase our knowledge of the fascinating, but complex system of the intestinal epithelium. This system is determined by an exchange of signals between the endoderm and the mesenchyme. Specific region differentiation, gut patterning and smooth muscle differentiation are directed not only by the BMP, Notch, Wnt and Hedgehog signalling pathways, but also by the *Hox, GATA, HNF, RXR, RORA, RAR, Apc* and *Cdx* genes, which orchestrate the cellular processes.

Bibliography

Bibliography

Afouda, B. A., A. Ciau-Uitz and R. Patient (2005). "GATA4, 5 and 6 mediate TGF beta maintenance of endodermal gene expression in Xenopus embryos." Development **132**(4): 763-774.

Allan, D., M. Houle, N. Bouchard, B. I. Meyer, P. Gruss and D. Lohnes (2001). "RAR gamma and Cdx1 interactions in vertebral patterning." Developmental Biology **240**(1): 46-60.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Research **25**(17): 3389-3402.

Amaya, E., P. A. Stein, T. J. Musci and M. W. Kirschner (1993). "Fgf Signaling in the Early Specification of Mesoderm in Xenopus." Development **118**(2): 477-487.

Aoki, K., Y. Tamai, S. Horiike, M. Oshima and M. M. Taketo (2003). "Colonic polyposis caused by mTOR-mediated chromosomal instability in Apc(+/Delta 716) Cdx(2+/-) compound mutant mice." Nature Genetics **35**(4): 323-330.

Aparicio, S., A. Morrison, A. Gould, J. Gilthorpe, C. Chaudhuri, P. Rigby, R. Krumlauf and S. Brenner (1995). "Detecting Conserved Regulatory Elements with the Model Genome of the Japanese Puffer Fish, Fugu Rubripes." Proceedings of the National Academy of Sciences of the United States of America **92**(5): 1684-1688.

Arnone, M. I., L. D. Bogarad, A. Collazo, C. V. Kirchhamer, R. A. Cameron, J. P. Rast, A. Gregorians and E. H. Davidson (1997). "Green Fluorescent Protein in the sea urchin: new experimental approaches to transcriptional regulatory analysis in embryos and larvae." Development **124**(22): 4649-4659.

Baeg, G. H., A. Matsumine, T. Kuroda, R. N. Bhattacharjee, I. Miyashiro, K. Toyoshima and T. Akiyama (1995). "The Tumor-Suppressor Gene-Product Apc

Blocks Cell-Cycle Progression from G(0)/G(1) to S-Phase." Embo Journal **14**(22): 5618-5625.

Barton, L. M., B. Gottgens, M. Gering, J. G. R. Gilbert, D. Grafham, J. Rogers, D. Bentley, R. Patient and A. R. Green (2001). "Regulation of the stem cell leukemia (SCL) gene: A tale of two fishes." Proceedings of the National Academy of Sciences of the United States of America **98**(12): 6747-6752.

Batlle, E., J. T. Henderson, H. Beghtel, M. M. W. van den Born, E. Sancho, G. Huis, J. Meeldijk, J. Robertson, M. van de Wetering, T. Pawson and H. Clevers (2002). "beta-catenin and TCF mediate cell positioning in the intestinal epithelium by controlling the expression of EphB/EphrinB." Cell **111**(2): 251-263.

Beck, F., K. Chawengsaksophak, J. Luckett, S. Giblett, J. Tucci, J. Brown, R. Poulsom, R. Jeffery and N. A. Wright (2003). "A study of regional gut endoderm potency by analysis of Cdx2 null mutant chimaeric mice." Developmental Biology **255**(2): 399-406.

Beck, F., K. Chawengsaksophak, P. Waring, R. J. Playford and J. B. Furness (1999). "Reprogramming of intestinal differentiation and intercalary regeneration in Cdx2 mutant mice." Proc Natl Acad Sci **96**(13): 7318-7323.

Beck, F., T. Erler, A. Russell and R. James (1995). "Expression of Cdx-2 in the mouse embryo and placenta: possible role in patterning of the extra-embrionic membranes." Dev Dyn **204**(3): 219-227.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick and D. Haussler (2004). "Ultraconserved elements in the human genome." Science **304**(5675): 1321-1325.

Beland, M., N. Pilon, M. Houle, K. Oh, J. R. Sylvestre, P. Prinos and D. Lohnes (2004). "Cdx1 autoregulation is governed by a novel Cdx1-LEF1 transcription complex." Molecular and Cellular Biology **24**(11): 5028-5038.

Berrueta, L., S. K. Kraeft, J. S. Tirnauer, S. C. Schuyler, L. B. Chen, D. E. Hill, D. Pellman and B. E. Bierer (1998). "The adenomatous polyposis coli-binding protein EB1 is associated with cytoplasmic and spindle microtubules." Proceedings of the National Academy of Sciences of the United States of America **95**(18): 10596-10601.

Blache, P., M. van de Wetering, I. Duluc, C. Domon, P. Berta, J. N. Freund, H. Clevers and P. Jay (2004). "SOX9 is an intestine crypt transcription factor, is regulated by the Wnt pathway, and represses the CDX2 and MUC2 genes." Journal of Cell Biology **166**(1): 37-47.

Boffelli, D., M. A. Nobrega and E. M. Rubin (2004). "Comparative genomics at the vertebrate extremes." Nature Reviews Genetics **5**(6): 456-465.

Boncinelli, E., A. Simeone, D. Acampora and F. Mavilio (1991). "Hox Gene Activation by Retinoic Acid." Trends in Genetics **7**(10): 329-334.

Bonner, C. A., S. K. Loftus and J. J. Wasmuth (1995). "Isolation, Characterization, and Precise Physical Localization of Human Cdx1, a Caudal-Type Homeobox Gene." Genomics **28**(2): 206-211.

Boudreau, F., E. Rings, H. Van Wering, G. P. Swain, S. D. Krasinski, E. R. Suh, R. J. Grand and P. G. Traber (2002). "HNF-1 alpha, GATA-4 and Cdx2 functionally interact to stimulate sucrase-isomaltase gene expression." Gastroenterology **122**(4): 161.

Boudreau, F., E. Rings, H. M. van Wering, R. K. Kim, G. P. Swain, S. D. Krasinski, J. Moffett, R. J. Grand, E. R. Suh and P. G. Traber (2002). "Hepatocyte nuclear factor-1 alpha, GATA-4, and caudal related homeodomain protein Cdx2 interact functionally to modulate intestinal gene transcription - Implication for the developmental regulation of the sucrose-isomaltase gene." Journal of Biological Chemistry **277**(35): 31909-31917.

Brantjes, H., N. Barker, J. van Es and H. Clevers (2002). "TCF: Lady justice casting the final verdict on the outcome of Wnt signalling." Biological Chemistry **383**(2): 255-261.

Brenner, S., G. Elgar, R. Sandford, A. Macrae, B. Venkatesh and S. Aparicio (1993). "Characterization of the Pufferfish (Fugu) Genome as a Compact Model Vertebrate Genome." Nature **366**(6452): 265-268.

Brooke, N. M., J. Garcia-Fernandez and P. W. H. Holland (1998). "The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster." Nature **392**(6679): 920-922.

Brudno, M., C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, E. D. Green, A. Sidow and S. Batzoglou (2003). "LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA." Genome Research **13**(4): 721-731.

Burglin, T. R., M. Finney, A. Coulson and G. Ruvkun (1989). "Caenorhabditis-Elegans Has Scores of Homeobox-Containing Genes." Nature **341**(6239): 239-243.

Chambon, P. (1996). "A decade of molecular biology of retinoic acid receptors." Faseb Journal **10**(9): 940-954.

Charite, J., W. de Graaff, D. Consten, M. J. Reijnen, J. Korving and J. Deschamps (1998). "Transducing positional information to the Hox genes: critical interaction of cdx gene products with position-sensitive regulatory elements." Development **125**(22): 4349-4358.

Chawengsaksophak, K., R. James, V. E. Hammond, F. Kontgen and F. Beck (1997). "Homeosis and intestinal tumours in Cdx2 mutant mice." Nature **386**(6620): 84-87.

Chung, C. T., S. L. Niemela and R. H. Miller (1989). "One-step preparation of competent Escherichia coli: transformation and storage of bacterial cells in the same solution." Proc Natl Acad Sci **86**(7): 2172-2175.

Clatworthy, J. P. and V. Subramanian (2001). "Stem cells and the regulation of proliferation, differentiation and patterning in the intestinal epithelium." Mech Dev **101**(1-2): 3-9.

Colnot, S., B. Romagnolo, M. Lambert, F. Cluzeaud, A. Porteu, A. Vandewalle, M. Thomasset, A. Kahn and C. Perret (1998). "Intestinal expression of the calbindin-D9K gene in transgenic mice - Requirement for a Cdx2-binding site in a distal activator region." Journal of Biological Chemistry **273**(48): 31939-31946.

Condie, B. G. and M. R. Capecchi (1993). "Mice Homozygous for a Targeted Disruption of Hoxd-3 (Hox-4.1) Exhibit Anterior Transformations of the First and Second Cervical-Vertebrae, the Atlas and the Axis." Development **119**(3): 579-595.

Consales, C. and M. I. Arnone (2002). "Functional characterization of Ets-binding sites in the sea urchin embryo: three base pair conversions redirect expression from mesoderm to ectoderm and endoderm." Gene **287**(1-2): 75-81.

Crawford, H. C., B. M. Fingleton, L. A. Rudolph-Owen, K. J. H. Goss, B. Rubinfeld, P. Polakis and L. M. Matrisian (1999). "The metalloproteinase matrilysin is a target of beta-catenin transactivation in intestinal tumors." Oncogene **18**(18): 2883-2891.

Davidson, A. J., P. Ernst, Y. Wang, M. P. S. Dekens, P. D. Kingsley, J. Palis, S. J. Korsmeyer, G. Q. Daley and L. I. Zon (2003). "cdx4 mutants fail to specify blood progenitors and can be rescued by multiple hox genes." Nature **425**(6955): 300-306.

Davidson, E. H. (2004). "Functional properties of the gene regulatory network for early sea urchin development." Faseb Journal **18**(5): A770-A770.

Davidson, E. H., D. R. McCay and L. Hood (2003). "Regulatory gene networks and the properties of the developmental process." Proceedings of the National Academy of Sciences of the United States of America **100**(4): 1475-1480.

Davidson, E. H., J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C. H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, A. G. Rust, Z. J. Pan, M. J. Schilstra, P. J. C. Clarke, M. I. Arnone, L. Rowen, R. A. Cameron, D. R. McClay, L. Hood and H. Bolouri (2002). "A genomic regulatory network for development." Science **295**(5560): 1669-1678.

Deka, J., P. Herter, M. Sprenger-Haussels, S. Koosch, D. Franz, K. M. Muller, C. Kuhnen, I. Hoffmann and O. Muller (1999). "The APC protein binds to A/T rich DNA sequences." Oncogene **18**(41): 5654-5661.

Driever, W., L. SolnicaKrezel, A. F. Schier, S. C. F. Neuhauss, J. Malicki, D. L. Stemple, D. Y. R. Stainier, F. Zwartkruis, S. Abdelilah, Z. Rangini, J. Belak and C. Boggs (1996). "A genetic screen for mutations affecting embryogenesis in zebrafish." Development **123**: 37-46.

Drummond, F., J. Sowden, K. Morrison and Y. H. Edwards (1996). "The caudal-type homeobox protein Cdx-2 binds to the colon promoter of the carbonic anhydrase 1 gene." European Journal of Biochemistry **236**(2): 670-681.

Dupe, V., N. B. Ghyselinck, O. Wendling, P. Chambon and M. Mark (1999). "Key roles of retinoic acid receptors alpha and beta in the patterning of the caudal hindbrain, pharyngeal arches and otocyst in the mouse." Development **126**(22): 5051-5059.

Duprey, P., K. Chowdhury, G. R. Dressler, R. Balling, D. Simon, J. L. Guenet and P. Gruss (1988). "A mouse gene homologous to the Drosophila gene caudal is expressed in epithelial cells from embryo intestine." Genes Dev **2**(12A): 1647-1654.

Dussault, I., D. Fawcett, A. Matthyssen, J. A. Bader and V. Giguere (1998). "Orphan nuclear receptor ROR alpha-deficient mice display the cerebellar defects of staggerer." Mechanisms of Development **70**(1-2): 147-153.

Edwards, Y. J. K., T. J. Carver, T. Vavouri, M. Frith, M. J. Bishop and G. Elgar (2003). "Theatre: a software tool for detailed comparative analysis and visualization of genomic sequence." Nucleic Acids Research **31**(13): 3510-3517.

Elgar, G., M. S. Clark, S. Meek, S. Smith, S. Warner, Y. J. K. Edwards, N. Bouchireb, A. Cottage, G. S. H. Yeo, Y. Umrania, G. Williams and S. Brenner (1999). "Generation and analysis of 25 Mb of genomic DNA from the pufferfish Fugu rubripes by sequence scanning." Genome Research **9**(10): 960-971.

Elgar, G., R. Sandford, S. Aparicio, A. Macrae, B. Venkatesh and S. Brenner (1996). "Small is beautiful: Comparative genomics with the pufferfish (Fugu rubripes)." Trends in Genetics **12**(4): 145-150.

Fearnhead, N. S., M. P. Britton and W. F. Bodmer (2001). "The ABC of APC." Human Molecular Genetics **10**(7): 721-733.

Flock, G. and D. J. Drucker (2002). "Pax-2 activates the proglucagon gene promoter but is not essential for proglucagon gene expression or development of proglucagon-producing cell lineages in the murine pancreas or intestine." Molecular Endocrinology **16**(10): 2349-2359.

Fodde, R. (2002). "The APC gene in colorectal cancer." Eur J Cancer **38**(7): 867-871.

Fodor, E., S. L. Weinrich, A. Meister, N. Mermod and W. J. Rutter (1991). "A Pancreatic Exocrine Cell Factor and Ap4 Bind Overlapping Sites in the Amylase 2a Enhancer." Biochemistry **30**(33): 8102-8108.

Gamer, L. W. and C. V. Wright (1993). "Murine Cdx-4 bears striking similarities to the Drosophila caudal gene in its homeodomain sequence and early expression pattern." Mech Dev **43**(1): 71-81.

Gaunt, S. J., D. Drage and A. Cockley (2003). "Vertebrate caudal gene expression gradients investigated by use of chick cdx-A/lacZ and mouse cdx-1/lacZ reporters in transgenic mouse embryos: evidence for an intron enhancer." Mechanisms of Development **120**(5): 573-586.

Gellner, K. and S. Brenner (1999). "Analysis of 148 kb of genomic DNA around the wnt1 locus of Fugu rubripes." Genome Research **9**(3): 251-258.

Gellon, G. and W. McGinnis (1998). "Shaping animal body plans in development and evolution by modulation of Hox expression patterns." Bioessays **20**(2): 116-125.

Gilbert, S. F. (2000). Developmental Biology. Printed U.S.A.

Gregorieff, A., R. Grosschedl and H. Clevers (2004). "Hindgut defects and transformation of the gastrointestinal tract in Tcf4(-/-)/Tcf1(-/-) embryos." Embo Journal **23**(8): 1825-1833.

Griffin, K. J. P., S. L. Amacher, C. B. Kimmel and D. Kimelman (1998). "Molecular identification of spadetail: regulation of zebrafish trunk and tail mesoderm formation by T-box genes." Development **125**(17): 3379-3388.

Groden, J., A. Thliveris, W. Samowitz, M. Carlson, L. Gelbert, H. Albertsen, G. Joslyn, J. Stevens, L. Spirio, M. Robertson, L. Sargeant, K. Krapcho, E. Wolff, R. Burt, J. P. Hughes, J. Warrington, J. McPherson, J. Wasmuth, D. Lepaslier, H. Abderrahim, D. Cohen, M. Leppert and R. White (1991). "Identification and Characterization of the Familial Adenomatous Polyposis-Coli Gene." Cell **66**(3): 589-600.

Haffter, P., M. Granato, M. Brand, M. C. Mullins, M. Hammerschmidt, D. A. Kane, J. Odenthal, F. J. M. vanEeden, Y. J. Jiang, C. P. Heisenberg, R. N. Kelsh, M. FurutaniSeiki, E. Vogelsang, D. Beuchle, U. Schach, C. Fabian and C. NussleinVolhard (1996). "The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio." Development **123**: 1-36.

Hart, M. J., R. de los Santos, I. N. Albert, B. Rubinfeld and P. Polakis (1998). "Downregulation of beta-catenin by human Axin and its association with the APC tumor suppressor, beta-catenin and GSK3 beta." Current Biology **8**(10): 573-581.

He, T. C., A. B. Sparks, C. Rago, H. Hermeking, L. Zawel, L. T. da Costa, P. J. Morin, B. Vogelstein and K. W. Kinzler (1998). "Identification of c-MYC as a target of the APC pathway." Science **281**(5382): 1509-1512.

Henderson, B. R. (2000). "Nuclear-cytoplasmic shuttling of APC regulates beta-catenin subcellular localization and turnover." Nature Cell Biology **2**(9): 653-660.

Her, G. M., Y. H. Yeh and J. L. Wu (2004). "Functional conserved elements mediate intestinal type fatty acid binding protein (I-FABP) expression in the gut epithelia of zebrafish larvae." Developmental Dynamics **230**(4): 734-742.

Hinoi, T., M. Tani, P. C. Lucas, K. Caca, R. L. Dunn, E. Macri, M. Loda, H. D. Appelman, K. R. Cho and E. R. Fearon (2001). "Loss of CDX2 expression and microsatellite instability are prominent features of large cell minimally differentiated carcinomas of the colon." American Journal of Pathology **159**(6): 2239-2248.

Holler M, Westin G, Jiricny J and S. W. (1988). "Sp1 transcription factor binds DNA and activates transcription even when the binding site is CpG methylated." Genes Dev. **2**(9): 1127-1135.

Horii, A., S. Nakatsuru, S. Ichii, H. Nagase and Y. Nakamura (1993). "Multiple Forms of the Apc Gene Transcripts and Their Tissue- Specific Expression." Human Molecular Genetics **2**(3): 283-287.

Houle, M., P. Prinos, A. Iulianella, N. Bouchard and D. Lohnes (2000). "Retinoic acid regulation of Cdx1: an indirect mechanism for retinoids and vertebral specification." Molecular and Cellular Biology **20**(17): 6579-6586.

Houle, M., J. R. Sylvestre and D. Lohnes (2003). "Retinoic acid regulates a subset of Cdx1 function in vivo." Development **130**(26): 6555-6567.

How, G. F., B. Venkatesh and S. Brenner (1996). "Conserved linkage between the puffer fish (Fugu rubripes) and human genes for platelet-derived growth factor receptor and macrophage colony-stimulating factor receptor." Genome Research **6**(12): 1185-1191.

Hsu, W., L. Zeng and F. Costantini (1999). "Identification of a domain of axin that binds to the serine/threonine protein phosphatase 2A and a self-binding domain." Journal of Biological Chemistry **274**(6): 3439-3445.

Hu, Y. L., J. Kazenwadel and R. James (1993). "Isolation and Characterization of the Murine Homeobox Gene Cdx- 1 Regulation of Expression in Intestinal Epithelial-Cells." Journal of Biological Chemistry **268**(36): 27214-27225.

Ikeya, M. and S. Takada (2001). "Wnt-3a is required for somite specification along the anteroposterior axis of the mouse embryo and for regulation of cdx-1 expression." Mechanisms of Development **103**(1-2): 27-33.

Isaacs, H. V., M. E. Pownall and J. M. W. Slack (1998). "Regulation of Hox gene expression and posterior development by the Xenopus caudal homologue Xcad3." Embo Journal **17**(12): 3413-3427.

Jacobsen, C. M., S. Mannisto, S. Porter-Tinge, E. Genova, H. Parviainen, M. Heikinheimo, Adameyko, II, S. G. Tevosian and D. B. Wilson (2005). "GATA-4: FOG interactions regulate gastric epithelial development in the mouse." Developmental Dynamics **234**(2): 355-362.

Jacobsen, C. M., N. Narita, M. Bielinska, A. J. Syder, J. I. Gordon and D. B. Wilson (2002). "Genetic mosaic analysis reveals that GATA-4 is required for proper differentiation of mouse gastric epithelium." Developmental Biology **241**(1): 34-46.

James, R., T. Erler and J. Kazenwadel (1994). "Structure of the Murine Homeobox Gene Cdx-2 - Expression in Embryonic and Adult Intestinal Epithelium." Journal of Biological Chemistry **269**(21): 15229-15237.

James, R. and J. Kazenwadel (1991). "Homeobox gene expression in the intestinal epithelium of adult mice." J Biol Chem **266**(5): 3246-3251.

Jette, C., P. W. Peterson, I. T. Sandoval, E. J. Manos, E. Hadley, C. M. Ireland and D. A. Jones (2004). "The tumor suppressor adenomatous polyposis coli and caudal related homeodomain protein regulate expression of retinol dehydrogenase L." Journal of Biological Chemistry **279**(33): 34397-34405.

Joly, J. S., M. Maury, C. Joly, P. Duprey, H. Boulekbache and H. Condamine (1992). "Expression of a Zebrafish Caudal Homeobox Gene Correlates with the Establishment of Posterior Cell Lineages at Gastrulation." Differentiation **50**(2): 75-87.

Jou, T. S., D. B. Stewart, J. Stappert, W. J. Nelson and J. A. Marrs (1995). "Genetic and Biochemical Dissection of Protein Linkages in the Cadherin-Catenin Complex." Proceedings of the National Academy of Sciences of the United States of America **92**(11): 5067-5071.

Kaestner, K. H., D. G. Silberg, P. G. Traber and G. Schutz (1997). "The mesenchymal winged helix transcription factor Fkh6 is required for the control of gastrointestinal proliferation and differentiation." Genes & Development **11**(12): 1583-1595.

Kane, K. F., M. J. S. Langman and G. R. Williams (1995). "1,25-Dihydroxyvitamin-D-3 and Retinoid-X Receptor Expression in Human Colorectal Neoplasms." Gut **36**(2): 255-258.

Kanki, J. P. and R. K. Ho (1997). "The development of the posterior body in zebrafish." Development **124**(4): 881-893.

Karagianni, N., M. C. Ly, S. Psarras, K. Chlichlia, V. Schirrmacher, F. Gounari and K. Khazaie (2005). "Novel adenomatous polyposis coli gene promoter is located 40 kb upstream of the initiating methionine." Genomics **85**(2): 231-237.

Karg H, Burger EH, B. A. Lyaruu DM and W. JH. (1997). "Spatiotemporal expression of the homeobox gene S8 during mouse tooth development." Arch Oral Biol. **42**(9): 625-631.

Kastner, P., M. Mark and P. Chambon (1995). "Nonsteroid nuclear receptors: what are genetic studies telling us about their role in real life." Cell **83**(6): 859-869.

Kaufman, M. H., R. M. Brune, R. A. Baldock, J. B. L. Bard and D. Davidson (1997). "Computer-aided 3-D reconstruction of serially sectioned mouse embryos: Its use in integrating anatomical organization." International Journal of Developmental Biology **41**(2): 223-233.

Kawasaki, Y., T. Senda, T. Ishidate, R. Koyama, T. Morishita, Y. Iwayama, O. Higuchi and T. Akiyama (2000). "Asef, a link between the tumor suppressor APC and G-protein signaling." Science **289**(5482): 1194-1197.

Kedinger, M., O. Lefebvre, I. Duluc, J. N. Freund and P. Simon-Assmann (1998). "Cellular and molecular partners involved in gut morphogenesis and differentiation." Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **353**(1370): 847-856.

Kelly, G. M., P. Greenstein, D. F. Erezyilmaz and R. T. Moon (1995). "Zebrafish Wnt8 and Wnt8b Share a Common Activity but Are Involved in Distinct Developmental Pathways." Development **121**(6): 1787-1799.

Kessel, M. and P. Gruss (1991). "Homeotic transformations of murine vertebrae and concomitant alteration of Hox codes induced by retinoic acid." Cell **67**(1): 89-104.

Kimmel, C. B., W. W. Ballard, S. R. Kimmel, B. Ullmann and T. F. Schilling (1995). "Stages of Embryonic-Development of the Zebrafish." Developmental Dynamics **203**(3): 253-310.

Kimura-Yoshida, C., K. Kitajima, I. Oda-Ishii, E. Tian, M. Suzuki, M. Yamamoto, T. Suzuki, M. Kobayashi, S. Aizawa and I. Matsuo (2004). "Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification." Development **131**(1): 57-71.

Kishida, S., H. Yamamoto, S. Ikeda, M. Kishida, I. Sakamoto, S. Koyama and A. Kikuchi (1998). "Communication - Axin, a negative regulator of the Wnt signaling pathway, directly interacts with adenomatous polyposis coli and regulates the stabilization of beta-catenin." Journal of Biological Chemistry **273**(18): 10823-10826.

Korinek, V., N. Barker, P. Moerer, E. van Donselaar, G. Huls, P. J. Peters and H. Clevers (1998). "Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4." Nature Genetics **19**(4): 379-383.

Korinek, V., N. Barker, P. J. Morin, D. vanWichen, R. deWeger, K. W. Kinzler, B. Vogelstein and H. Clevers (1997). "Constitutive transcriptional activation by a beta-catenin-Tcf complex in APC(-/-) colon carcinoma." Science **275**(5307): 1784-1787.

Krasinski, S. D., H. M. Van Wering, M. R. Tannemaat and R. J. Grand (2001). "Differential activation of intestinal gene promoters: functional interactions between GATA-5 and HNF-1 alpha." American Journal of Physiology-Gastrointestinal and Liver Physiology **281**(1): G69-G84.

Larsson, L. I., L. St-Onge, D. M. Hougaard, B. Sosa-Pineda and P. Gruss (1998). "Pax 4 and 6 regulate gastrointestinal endocrine cell development." Mechanisms of Development **79**(1-2): 153-159.

Lee, S. Y., B. P. Nagy, A. R. Brooks, D. M. Wang, B. Paulweber and B. LevyWilson (1996). "Members of the caudal family of homeodomain proteins repress transcription from the human apolipoprotein B promoter in intestinal cells." Journal of Biological Chemistry **271**(2): 707-718.

Lee, Y. J., B. Swencki, S. Shoichet and R. A. Shivdasani (1999). "A possible role for the high mobility group box transcription factor Tcf-4 in vertebrate gut epithelial cell differentiation." Journal of Biological Chemistry **274**(3): 1566-1572.

Lekven, A. C., C. J. Thorpe, J. S. Waxman and R. T. Moon (2001). "Zebrafish wnt8 encodes two wnt8 proteins on a bicistronic transcript and is required for mesoderm and neurectoderm patterning." Developmental Cell **1**(1): 103-114.

Lickert, H., C. Domon, G. Huls, C. Wehrle, I. Duluc, H. Clevers, B. I. Meyer, J. N. Freund and R. Kemler (2000). "Wnt/(beta)-catenin signaling regulates the expression of the homeobox gene Cdx1 in the embryonic intestine." Development **127**(17): 3805-3813.

Lickert, H. and R. Kemler (2002). "Functional analysis of cis-regulatory elements controlling initiation and maintenance of early Cdx1 gene expression in the mouse." Developmental Dynamics **225**(2): 216-220.

Lorentz, O., A. Cadoret, I. Duluc, J. Capeau, C. Gespach, G. Cherqui and J. N. Freund (1999). "Downregulation of the colon tumour-suppressor homeobox gene Cdx-2 by oncogenic ras." Oncogene **18**(1): 87-92.

Lorentz, O., I. Duluc, A. D. Arcangelis, P. Simon-Assmann, M. Kedinger and J. N. Freund (1997). "Key role of the Cdx2 homeobox gene in extracellular matrix-mediated." J Cell Biol **139**(6): 1553-1565.

Mahmoud, N. N., S. K. Boolbol, R. T. Bilinski, C. Martucci, A. Chadburn and M. M. Bertagnolli (1997). "Apc gene mutation is associated with a dominant-negative effect upon intestinal cell migration." Cancer Research **57**(22): 5045-5050.

Mallo, G. V., P. Soubeyran, J. C. Lissitzky, F. Andre, C. Farnarier, J. Marvaldi, J. C. Dagorn and J. L. Iovanna (1998). "Expression of the Cdx1 and Cdx2 homeotic genes leads to reduced malignancy in colon cancer-derived cells." Journal of Biological Chemistry **273**(22): 14030-14036.

Mangelsdorf, D. J., C. Thummel, M. Beato, P. Herrlich, G. Schutz, K. Umesono, B. Blumberg, P. Kastner, M. Mark and P. Chambon (1995). "The nuclear receptor superfamily: the second decade." Cell **83**(6): 835-839.

Marshall, H., A. Morrison, M. Studer, H. Popperl and R. Krumlauf (1996). "Retinoids and Hox genes." Faseb Journal **10**(9): 969-978.

Marshall, H., M. Studer, H. Popperl, S. Aparicio, A. Kuroiwa, S. Brenner and R. Krumlauf (1994). "A Conserved Retinoic Acid Response Element Required for Early Expression of the Homeobox Gene Hoxb-1." Nature **370**(6490): 567-571.

Merika, M. and S. H. Orkin (1995). "Functional Synergy and Physical Interactions of the Erythroid Transcription Factor Gata-1 with the Kruppel Family Proteins Sp1 and Eklf." Molecular and Cellular Biology **15**(5): 2437-2447.

Meyer, B. I. and P. Gruss (1993). "Mouse Cdx-1 expression during gastrulation." Development **117**(1): 191-203.

Mlodzik, M., A. Fjose and W. J. Gehring (1985). "Isolation of Caudal, a Drosophila Homeo Box-Containing Gene with Maternal Expression, Whose Transcripts Form a Concentration Gradient at the Pre-Blastoderm Stage." Embo Journal **4**(11): 2961-2969.

Moore-Scott, B. A. and N. R. Manley (2005). "Differential expression of Sonic hedgehog along the anterior-posterior axis regulates patterning of pharyngeal pouch endoderm and pharyngeal endoderrn-derived organs." Developmental Biology **278**(2): 323-335.

Morin, P. J., A. B. Sparks, V. Korinek, N. Barker, H. Clevers, B. Vogelstein and K. W. Kinzler (1997). "Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC." Science **275**(5307): 1787-1790.

Morrison, E. E., B. N. Wardleworth, J. M. Askham, A. F. Markham and D. M. Meredith (1998). "EB1, a protein which interacts with the APC tumour suppressor, is associated with the microtubule cytoskeleton throughout the cell cycle." Oncogene **17**(26): 3471-3477.

Moucadel, V., P. Soubeyran, S. Vasseur, N. J. Dusetti, J. C. Dagorn and J. L. Iovanna (2001). "Cdx1 promotes cellular growth of epithelial intestinal cells through induction of the secretory protein PAP I." European Journal of Cell Biology **80**(2): 156-163.

Moucadel, V., M. S. Totaro, C. D. Dell, P. Soubeyran, J. C. Dagorn, J. N. Freund and J. L. Iovanna (2002). "The homeobox gene Cdx1 belongs to the p53-p21(WAF)-Bcl-2 network in intestinal epithelial cells." Biochemical and Biophysical Research Communications **297**(3): 607-615.

Muller, F., P. Blader and U. Strahle (2002). "Search for enhancers: teleost models in comparative genomic and transgenic analysis of cis regulatory elements." Bioessays **24**(6): 564-572.

Munemitsu, S., B. Souza, O. Muller, I. Albert, B. Rubinfeld and P. Polakis (1994). "The Apc Gene-Product Associates with Microtubules in-Vivo and Promotes Their Assembly in-Vitro." Cancer Research **54**(14): 3676-3681.

Nagafuchi, A. and M. Takeichi (1988). "Cell Binding Function of E-Cadherin Is Regulated by the Cytoplasmic Domain." Embo Journal **7**(12): 3679-3684.

Nakagawa, H., Y. Murata, K. Koyama, A. Fujiyama, Y. Miyoshi, M. Monden, T. Akiyama and Y. Nakamura (1998). "Identification of a brain-specific APC homologue, APCL, and its interaction with beta-catenin." Cancer Research **58**(22): 5176-5181.

Neufeld, K. L. and R. L. White (1997). "Nuclear and cytoplasmic localizations of the adenomatous polyposis coli protein." Proceedings of the National Academy of Sciences of the United States of America **94**(7): 3034-3039.

Niessing, D., S. Blanke and H. Jackle (2002). "Bicoid associates with the 5 '-cap-bound complex of caudal mRNA and represses translation." Genes & Development **16**(19): 2576-2582.

Oates, A. C., P. Wollberg, S. J. Pratt, B. H. Paw, S. L. Johnson, R. K. Ho, J. H. Postlethwait, L. I. Zon and A. F. Wilks (1999). "Zebrafish stat3 is expressed in restricted tissues during embryogenesis and stat1 rescues cytokine signaling in a STAT1-deficient human cell line." Developmental Dynamics **215**(4): 352-370.

Ozawa, M., H. Baribault and R. Kemler (1989). "The Cytoplasmic Domain of the Cell-Adhesion Molecule Uvomorulin Associates with 3 Independent Proteins Structurally Related in Different Species." Embo Journal **8**(6): 1711-1717.

Patient, R. K. and J. D. McGhee (2002). "The GATA family (vertebrates and invertebrates)." Current Opinion in Genetics & Development **12**(4): 416-422.

Pennacchio, L. A. and E. M. Rubin (2001). "Genomic strategies to identify mammalian regulatory sequences." Nature Reviews Genetics **2**(2): 100-109.

Pillemer, G., M. Epstein, B. Blumberg, J. K. Yisraeli, E. M. De Robertis, H. Steinbeisser and A. Fainsod (1998). "Nested expression and sequential downregulation of the Xenopus caudal genes along the anterior-posterior axis." Mechanisms of Development **71**(1-2): 193-196.

Polakis, P. (1997). "The adenomatous polyposis coli (APC) tumor suppressor." Biochimica Et Biophysica Acta-Reviews on Cancer **1332**(3): F127-F147.

Polakis, P. (2000). "Wnt signaling and cancer." Genes & Development **14**(15): 1837-1851.

Potten, C. S. and M. Loeffler (1990). "Stem-Cells - Attributes, Cycles, Spirals, Pitfalls and Uncertainties - Lessons for and from the Crypt." Development **110**(4): 1001-1020.

Pownall, M. E., A. S. Tucker, J. M. W. Slack and H. V. Isaacs (1996). "eFGF, Xcad3 and Hox genes form a molecular pathway that establishes the anteroposterior axis in Xenopus." Development **122**(12): 3881-3892.

Prinos, P., S. Joseph, K. Oh, B. I. Meyer, P. Gruss and D. Lohnes (2001). "Multiple pathways governing Cdx1 expression during murine development." Dev Biol **239**(2): 257-269.

Quandt, K., K. Frech, H. Karas, E. Wingender and T. Werner (1995). "MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data." Nucleic Acids Res. **23**(23): 4878-4884.

Rankin, E. B., W. Xu, D. G. Silberg and E. Suh (2004). "Putative intestine-specific enhancers located in 5 ' sequence of the CDX1 gene regulate CDX1 expression in the intestine." American Journal of Physiology-Gastrointestinal and Liver Physiology **286**(5): G872-G880.

Rawat, V. P. S., M. Cusan, A. Deshpande, W. Hiddemann, L. Quintanilla-Martinez, R. K. Humphries, S. K. Bohlander, M. Feuring-Buske and C. Buske (2004). "Ectopic expression of the homeobox gene Cdx2 is the transforming event in a mouse model of t(12;13)(p13;q12) acute myeloid leukemia." Proceedings of the National Academy of Sciences of the United States of America **101**(3): 817-822.

Rice, P., I. Longden and A. Bleasby (2000). "EMBOSS: The European molecular biology open software suite." Trends in Genetics **16**(6): 276-277.

RiveraPomar, R., D. Niessing, U. SchmidtOtt, W. J. Gehring and H. Jackle (1996). "RNA binding and translational suppression by bicoid." Nature **379**(6567): 746-749.

Rosin-Arbesfeld, R., F. Townsley and M. Bienz (2000). "The APC tumour suppressor has a nuclear export function." Nature **406**(6799): 1009-1012.

Rothenberg, E. V. (2001). "Mapping of complex regulatory elements by pufferfish/zebrafish transgenesis." Proceedings of the National Academy of Sciences of the United States of America **98**(12): 6540-6542.

Rowitch, D. H., Y. Echelard, P. S. Danielian, K. Gellner, S. Brenner and A. P. McMahon (1998). "Identification of an evolutionarily conserved 110 base-pair cis-acting regulatory sequence that governs Wnt-1expression in the murine neural plate." Development **125**(14): 2735-2746.

Rubinfeld, B., I. Albert, E. Porfiri, S. Munemitsu and P. Polakis (1997). "Loss of beta-catenin regulation by the APC tumor suppressor protein correlates with loss of structure due to common somatic mutations of the gene." Cancer Research **57**(20): 4624-4630.

Rubinfeld, B., B. Souza, I. Albert, O. Muller, S. H. Chamberlain, F. R. Masiarz, S. Munemitsu and P. Polakis (1993). "Association of the Apc Gene-Product with Beta-Catenin." Science **262**(5140): 1731-1734.

Ruvinsky, I., A. C. Oates, L. M. Silver and R. K. Ho (2000). "The evolution of paired appendages in vertebrates: T-box genes in the zebrafish." Development Genes and Evolution **210**(2): 82-91.

Sanbrook, J., Fritsch, E. F., and Maniatis, T. (1989). Molecular Cloning. A laboratory manual. U.S.A.

Santoro, I. M. and J. Groden (1997). "Alternative splicing of the APC gene and its association with terminal differentiation." Cancer Res **57**(3): 488-494.

Seeling, J. M., J. R. Miller, R. Gil, R. T. Moon, R. White and D. M. Virshup (1999). "Regulation of beta-catenin signaling by the B56 subunit of protein phosphatase 2A." Science **283**(5410): 2089-2091.

Seipel, K., O. Georgiev and W. Schaffner (1992). "Different Activation Domains Stimulate Transcription from Remote (Enhancer) and Proximal (Promoter) Positions." Embo Journal **11**(13): 4961-4968.

Seligson, D., S. Horvath, S. Huerta-Yepez, S. Hanna, H. Garban, A. Roberts, T. Shi, X. L. Liu, D. Chia, L. Goodglick and B. Bonavida (2005). "Expression of transcription factor Yin Yang 1 in prostate cancer." International Journal of Oncology **27**(1): 131-141.

Senda, T., S. Iino, K. Matsushita, A. Matsumine, S. Kobayashi and T. Akiyama (1998). "Localization of the adenomatous polyposis coli tumour suppressor protein in the mouse central nervous system." Neuroscience **83**(3): 857-866.

Shtutman, M., J. Zhurinsky, I. Simcha, C. Albanese, M. D'Amico, R. Pestell and A. Ben-Ze'ev (1999). "The cyclin D1 gene is a target of the beta-catenin/LEF-1 pathway." Proceedings of the National Academy of Sciences of the United States of America **96**(10): 5522-5527.

Silberg, D. G., G. P. Swain, E. R. Suh and P. G. Traber (2000). "Cdx1 and cdx2 expression during intestinal development." Gastroenterology **119**(4): 961-971.

Singh, S., R. Poulsom, N. A. Wright, M. C. Sheppard and M. J. S. Langman (1997). "Differential expression of oestrogen receptor and oestrogen inducible genes in gastric mucosa and cancer." Gut **40**(4): 516-520.

Smith, K. J., K. A. Johnson, T. M. Bryan, D. E. Hill, S. Markowitz, J. K. V. Willson, C. Parasekva, G. M. Petersen, S. R. Hamilton, B. Vogelstein and K. W. Kinzler (1993). "The Apc Gene-Product in Normal and Tumor-Cells." Proceedings of the National Academy of Sciences of the United States of America **90**(7): 2846-2850.

Smits, R., M. F. Kielman, C. Breukel, C. Zurcher, K. Neufeld, S. Jagmohan-Changur, N. Hofland, J. van Dijk, R. White, W. Edelmann, R. Kucherlapati, P. M. Khan and R. Fodde (1999). "Apc1638T: a mouse model delineating critical domains of

the adenomatous polyposis coli protein involved in tumorigenesis and development." Genes & Development **13**(10): 1309-1321.

Soubeyran, P., F. Andre, J. C. Lissitzky, G. V. Mallo, V. Moucadel, M. Roccabianca, H. Rechreche, J. Marvaldi, I. Dikic, J. C. Dagorn and J. L. Iovanna (1999). "Cdx1 promotes differentiation in a rat intestinal epithelial cell line." Gastroenterology **117**(6): 1326-1338.

Sprague, J., E. Doerry, S. Douglas and M. and Westerfield (2001). "The Zebrafish Information Network (ZFIN): a resource for genetic, genomic and developmental research." Nucleic Acids Research **29**: 87-90.

Stappenbeck, T. S., L. V. Hooper and J. I. Gordon (2002). "Developmental regulation of intestinal angiogenesis by indigenous microbes via Paneth cells." Proceedings of the National Academy of Sciences of the United States of America **99**(24): 15451-15455.

Struhl, G. and R. A. H. White (1985). "Regulation of the Ultrabithorax Gene of Drosophila by Other Bithorax Complex Genes." Cell **43**(2): 507-519.

Su, L. K., K. A. Johnson, K. J. Smith, D. E. Hill, B. Vogelstein and K. W. Kinzler (1993). "Association between Wild-Type and Mutant Apc Gene-Products." Cancer Research **53**(12): 2728-2731.

Subramanian, V., B. Meyer and G. S. Evans (1998). "The murine Cdx1 gene product localises to the proliferative compartment in the developing and regenerating intestinal epithelium." Differentiation **64**(1): 11-18.

Subramanian, V., B. I. Meyer and P. Gruss (1995). "Disruption of the murine homeobox gene Cdx1 affects axial skeletal identities by altering the mesodermal expression domains of Hox genes." Cell **83**(4): 641-653.

Suh, E., L. L. Chen, J. Taylor and P. G. Traber (1994). "A Homeodomain Protein Related to Caudal Regulates Intestine- Specific Gene-Transcription." Molecular and Cellular Biology **14**(11): 7340-7351.

Suh, E. and T. P. G. (1996). "An intestine-specific homeobox gene regulates proliferation and." Mol Cell Biol **16**(2): 619-625.

Szpirer, C., M. Riviere, R. Cortese, T. Nakamura, M. Q. Islam, G. Levan and J. Szpirer (1992). "Chromosomal Localization in Man and Rat of the Genes Encoding the Liver-Enriched Transcription Factors-C/Ebp, Dbp, and Hnf1/Lfb-1 (Cebp, Dbp, and Transcription Factor-I, Tcf1, Respectively) and of the Hepatocyte Growth-Factor Scatter Factor Gene (Hgf)." Genomics **13**(2): 293-300.

Taylor, J. K., T. Levy, E. R. Suh and P. G. Traber (1997). "Activation of enhancer elements by the homeobox gene Cdx2 is cell line specific." Nucleic Acids Research **25**(12): 2293-2300.

Theodosiou, N. A. and C. J. Tabin (2005). "Sox9 and Nkx2.5 determine the pyloric sphincter epithelium under the control of BMP signaling." Developmental Biology **279**(2): 481-490.

Thliveris, A., W. Samowitz, N. Matsunami, J. Groden and R. White (1994). "Demonstration of Promoter Activity and Alternative Splicing in the Region 5' to Exon-1 of the Apc Gene." Cancer Research **54**(11): 2991-2995.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position- Specific Gap Penalties and Weight Matrix Choice." Nucleic Acids Research **22**(22): 4673-4680.

Tirnauer, J. S. and B. E. Bierer (2000). "EB1 proteins regulate microtubule dynamics, cell polarity, and chromosome stability." Journal of Cell Biology **149**(4): 761-766.

Troelsen, J. T., C. Mitchelmore, N. Spodsberg, A. M. Jensen, O. Noren and H. Sjostrom (1997). "Regulation of lactase-phlorizin hydrolase gene expression by

the caudal-related homoeodomain protein Cdx-2." Biochemical Journal **322**: 833-838.

Tronche, F. and M. Yaniv (1992). "Hnf1, a Homeoprotein Member of the Hepatic Transcription Regulatory Network." Bioessays **14**(9): 579-587.

van den Akker, E., S. Forlani, K. Chawengsaksophak, W. de Graaff, F. Beck, B. I. Meyer and J. Deschamps (2002). "Cdx1 and Cdx2 have overlapping functions in anteroposterior patterning and posterior axis elongation." Development **129**(9): 21822-2193.

van der Heyden, M. A., P. A. O. Weernink, B. A. van Oirschot, P. en Henegouwen, J. Boonstra and G. Rijksen (1997). "Epidermal growth factor-induced activation and translocation of c-Src to the cytoskeleton depends on the actin binding domain of the ECF-receptor." Biochimica Et Biophysica Acta-Molecular Cell Research **1359**(3): 211-221.

van Es, J. H., C. Kirkpatrick, M. van de Wetering, M. Molenaar, A. Miles, J. Kuipers, O. Destree, M. Peifer and H. Clevers (1999). "Identification of APC2, a homologue of the adenomatous polyposis coli tumour suppressor." Current Biology **9**(2): 105-108.

van Heel, D. A., I. A. Udalova, A. P. De Silva, D. P. McGovern, Y. Kinouchi, J. Hull, N. J. Lench, L. R. Cardon, A. H. Carey, D. P. Jewell and D. Kwiatkowski (2002). "Inflammatory bowel disease is associated with a TNF polymorphism that affects an interaction between the OCT1 and NF-kappa B transcription factors." Human Molecular Genetics **11**(11): 1281-1289.

Veeman, M. T., D. C. Slusarski, A. Kaykas, S. H. Louie and R. T. Moon (2003). "Zebrafish prickle, a modulator of noncanonical Wnt/Fz signaling, regulates gastrulation movements." Current Biology **13**(8): 680-685.

Venkatesh, B., S. L. SiHoe, D. Murphy and S. Brenner (1997). "Transgenic rats reveal functional conservation of regulatory controls between the Fugu isotocin

and rat oxytocin genes." Proceedings of the National Academy of Sciences of the United States of America **94**(23): 12462-12466.

Venkatesh, B., B. H. Tay, G. Elgar and S. Brenner (1996). "Isolation, characterization and evolution of nine pufferfish (Fugu rubripes) actin genes." Journal of Molecular Biology **259**(4): 655-665.

Wallace, K. N., S. Akhter, E. M. Smith, K. Lorent and M. Pack (2005). "Intestinal growth and differentiation in zebrafish." Mechanisms of Development **122**(2): 157-173.

Wallace, K. N. and M. Pack (2003). "Unique and conserved aspects of gut development in zebrafish." Developmental Biology **255**(1): 12-29.

Wedgwood, S., W. K. Lam, K. M. Pinchin, A. F. Markham, E. J. Cartwright and P. L. Coletta (2000). "Characterization of a brain-selective transcript of the Adenomatous polyposis coli tumor suppressor gene." Mamm Genome **11**(12): 1150-1153.

Willert, K., M. Brink, A. Wodarz, H. Varmus and R. Nusse (1997). "Casein kinase 2 associates with and phosphorylates dishevelled." Embo Journal **16**(11): 3089-3096.

Wolpert, L., R. Beddington, T. Jessell, P. Lawrence, E. Meyerowitz and J. Smith (2002). Principles of Development. Oxford, Oxford University Press.

Wong, M. H., M. L. Hermiston, A. J. Syder and J. I. Gordon (1996). "Forced expression of the tumor suppressor adenomatosis polyposis coli protein induces disordered cell migration in the intestinal epithelium." Proceedings of the National Academy of Sciences of the United States of America **93**(18): 9588-9593.

Woolfe, A., M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, S. F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y. J. K. Edwards, J. E. Cooke and G. Elgar (2005). "Highly conserved non-coding

sequences are associated with vertebrate development." Plos Biology **3**(1): 116-130.

Wu, L. H. and J. A. Lengyel (1998). "Role of caudal in hindgut specification and gastrulation suggests homology between Drosophila amnioproctodeal invagination and vertebrate blastopore." Development **125**(13): 2433-2442.

Yang, N., R. Schule, D. J. Mangelsdorf and R. M. Evans (1991). "Characterization of DNA binding and retinoic acid binding properties of retinoic acid receptor." Proc Natl Acad Sci **88**(9): 3559-3563.

Ye, H. G., T. F. Kelly, U. Samadani, L. Lim, S. Rubio, D. G. Overdier, K. A. Roebuck and R. H. Costa (1997). "Hepatocyte nuclear factor 3/fork head homolog 11 is expressed in proliferating epithelial and mesenchymal cells of embryonic and adult tissues." Molecular and Cellular Biology **17**(3): 1626-1641.

Zhang, F., H. Popperl, A. Morrison, E. N. Kovacs, V. Prideaux, L. Schwarz, R. Krumlauf, J. Rossant and M. S. Featherstone (1997). "Elements both 5' and 3' to the murine Hoxd4 gene establish anterior borders of expression in mesoderm and neurectoderm." Mechanisms of Development **67**(1): 49-58.

APPENDIX

Appendix, Section 1A

Cdx1 nucleotide alignment

```
5'UTR Cdx1
mmCdx15UTR ------GCAGTCGCTGGTCGTCGGGGCGGCTCGCTCGGGCGCGGCGGCGCCAGGGCCCAG 54
hsCdx15UTR AGGTGAGCGGTTGCTCGTCGTCGGGGCGGC-----CGGCAGCGGCGGCTCCAGGGCCCAG 55
frCdx15UTR GAATGCTGGGAGACCTGACATGATCCCAGCT----CTTTGATTTTTGTCCAGTGTCACAT 56
                  *   *  * * *     * **     *          *  *   * * **

mmCdx15UTR CATGCGCGGGGGACCCTGCGGTCACCATG 83
hsCdx15UTR CATGCGCGGGGGACCCCGCGGCCACCATG 84
frCdx15UTR TTTACGCACCAGGACGCAGGAA----ATG 81
            * ***    *  *    *   *  ***

2nd exon Cdx1
mmCdx1 GTAAGACCCGAACCAAGGACAAGTACCGTGTGGTCTACACAGACCACCAACGCCTAGAGC 60
hsCdx1 GTAAGACTCGGACCAAGGACAAGTACCGCGTGGTCTACACCGACCACCAACGCCTGGAGC 60
frCdx1 GGAAGACTCGCACTAAGGACAAGTACAGGGTGGTGTACACTGACAAACAGCGGATGGAGC 60
       * ***** ** ** ************ * ***** ***** *** * ** **  * ****

mmCdx1 TGGAAAAGGAGTTTCACTACAGCCGGTACATCACTATCCGGCGCAAGTCCGAGCTGGCTG 120
hsCdx1 TGGAGAAGGAGTTTCATTACAGCCGTTACATCACAATCCGGCGGAAATCAGAGCTGGCTG 120
frCdx1 TGGAGAGGGAGTTCCAGAGCAACCGCTACATCACCATGCGCAGGAAAGCAGAGCTGTCGA 120
       **** * ****** **   ** *** ******** ** **  * **  * ****** *

mmCdx1 CTAACCTGGGGCTCACAGAGCGGCAG 146
hsCdx1 CCAATCTGGGGCTCACTGAACGGCAG 146
frCdx1 TCACGCTGGGCCTCTCAGAGAGACAG 146
         *  ***** *** * **  * ***
```

```
3rd exon Cdx1
mmCdx1 GTAAAGATCTGGTTCCAGAACCGCCGGGCCAAGGAGCGCAAAGTAAACAAGAAGAAACAG 60
hsCdx1 GTGAAGATCTGGTTCCAAAACCGGCGGGCAAAGGAGCGCAAAGTGAACAAGAAGAAACAG 60
frCdx1 ATAAAAATATGGTTTCAGAACAGGCGTGCCAAAGAAAGAAAAATGAACCGGAAGAAGCTG 60
        * ** ** ***** ** *** * ** ** ** **  * *** * ***  ****** * *

mmCdx1 CAGCA---GCAGCAGCCCCTGCCTCCCACACAGCTGCCCCTGCCCCTGGATGGCACTCCC 117
hsCdx1 CAGCA---GCAACAGCCCC---------CACAGCCGCCGATGGCCCACGACATCACGGCC 108
frCdx1 CAGCATTCGCAGCAGGCCTC--------TACGACCACCCCCGCCTCCCCGGGCCTCGCTG 112
       *****   *** *** **  **  *  **   * * *       * *

mmCdx1 ACACCATCAGGGCCACCCCTAGGAAGTCTATGCCCTACTAATGCTGGCCTTCTGGGCACC 177
hsCdx1 ACCCCAGCCGGGCCATCCCTGGGGGGCCTGTGTCCCAGCAACACCAGCCTCCTGGCCACC 168
frCdx1 AGCCCGTGGAAGCCCACCCC----GGCATGAGCCCCAACGGCTTT---TTTTCAGACACC 165
       *  **      ***  ***      *  *  * ** *   *    * ****

mmCdx1 CCCTCCCCAGTGCCCGTCAAGGAGGAGTTTCTACCCTAG 216
hsCdx1 TCCTCTCCAATGCCTGTGAAAGAGGAGTTTCTGCCATAG 207
frCdx1 CTGTCAAAAGAATAT----------------------- 180
          **   *
```

Appendix, Section 1B

Cdx2 nucleotide alignment

```
1st exon Cdx2
mmCdx2 ATGTACGTGAGCTACCTTCTGGACAAGGACGTGAGCATGTATCCTAGCTCCGTGCGCCAC 60
hsCdx2 ATGTACGTGAGCTACCTCCTGGACAAGGACGTGAGCATGTACCCTAGCTCCGTGCGCCAC 60
frCdx2 ATGTATGTATGTCTATATGTTAATATGAGGACAGTTCTGTCATATTGTTGC-TGTGCGTG 59
       ***** **  *        *  * * *          ***    * * * * ** **

mmCdx2 TCCGGCGGCCTGAACCTGGCTCCGCAGAACTTTGTCAGTCCTCCGCAGTACCCGGACTAC 120
hsCdx2 TCTGGCGGCCTCAACCTGGCGCCGCAGAACTTCGTCAGCCCCCCGCAGTACCCGGACTAC 120
frCdx2 TCTAGTTG---------------------------------------------------- 67
```

```
      **  *  *

mmCdx2 GGTGGTTACCACGTGGCGGCCGCGGCGGCTGCTACGGCGAACTTGGACAGCGCTCAGTCC 180
hsCdx2 GGCGGTTACCACGTGGCGGCCGCAGCTGCAGCGGCAGCGAACTTGGACAGCGCGCAGTCC 180
frCdx2 ------------------------------------------------------------


mmCdx2 CCAGGGCCATCCTGGCCCACCGCGTACGGCGCCCCTCTCCGCGAGGACTGGAATGGCTAC 240
hsCdx2 CCGGGGCCATCCTGGCCGGCAGCGTATGGCGCCCCACTCCGGGAGGACTGGAATGGCTAC 240
frCdx2 ------------------------------------------------------------


mmCdx2 GCACCCGGGGGCGCTGCGGCAGCC---AACGCGGTAGCCCACGGTCTCAATGGTGGCTCC 297
hsCdx2 GCGCCCGGAGGCGCCGCGGCCGCCGCCAACGCCGTGGCTCACGGCCTCAACGGTGGCTCC 300
frCdx2 ------------------------------------------------------------


mmCdx2 CCGGCCGCCGCTATGGGCTACAGCAGCCCCGCCGAATACCACGCGCACCATCACCCGCAT 357
hsCdx2 CCGGCCGCAGCCATGGGCTACAGCAGCCCCGCAGACTACCATCCGCACCACCACCCGCAT 360
frCdx2 ------------------------------------------------------------


mmCdx2 CATCACCCGCACCATCCGGCCGCCTCGCCGTCCTGCGCCTCCGGCTTGCTGCAGACGCTC 417
hsCdx2 CACCACCCGCACCACCCGGCCGCCGCGCCTTCCTGCGCTTCTGGGCTGCTGCAAACGCTC 420
frCdx2 ------------------------------------------------------------


mmCdx2 AACCTCGGCCCCCCGGGGCCCGCAGCCACCGCCGCCGCCGAACAGCTGTCCCCCAGCGGC 477
hsCdx2 AACCCCGGCCCTCCTGGGCCCGCCGCCACCGCTGCCGCCGAGCAGCTGTCTCCCGGCGGC 480
frCdx2 ------------------------------------------------------------


mmCdx2 CAGCGGCGAAACCTGTGCGAGTGGATGCGGAAGCCCGCGCAGCAGTCCCTAGGAAGCCAAG 538
hsCdx2 CAGCGGCGGAACCTGTGCGAGTGGATGCGGAAGCCGGCGCAGCAGTCCCTCGGCAGCCAAG 541
frCdx2 ------------------------------------------------------------
```

```
2nd exon Cdx2
mmCdx2 TGAAAACCAGGACAAAAGACAAATACCGGGTGGTGTACACAGACCATCAGCGGCTGGAGC 60
hsCdx2 TGAAAACCAGGACGAAAGACAAATATCGAGTGGTGTACACGGACCACCAGCGGCTGGAGC 60
frCdx2 GAAAAACGCGGACTAAAGACAAGTACCGGGTGGTTTACACCGACCACCAGCGGCTGGAGC 60
         *****  **** ******** ** ** ***** ***** ***** *************

mmCdx2 TGGAGAAGGAGTTTCACTTTAGTCGATACATCACCATCAGGAGGAAAAGTGAGCTGGCTG 120
hsCdx2 TGGAGAAGGAGTTTCACTACAGTCGCTACATCACCATCCGGAGGAAAGCCGAGCTAGCCG 120
frCdx2 TGGAAAAGGAGTTTCACTACAGCAAGTACATCACCATCAGGAGGAAATCGGAGCTGGCCA 120
       **** *************  **    ************ ********   ***** **

mmCdx2 CCACACTTGGGCTCTCCGAGAGGCAG 146
hsCdx2 CCACGCTGGGGCTCTCTGAGAGGCAG 146
frCdx2 CAGCCCTCAGCCTATCAGAGCGACAG 146
       *  * **  * ** ** *** * ***

3rd exon Cdx2
mmCdx2 GTTAAAATTTGGTTTCAGAACCGCAGAGCCAAGGAGAGGAAAATCAA---GAAGAAGCAG 57
hsCdx2 GTTAAAATCTGGTTTCAGAACCGCAGAGCAAAGGAGAGGAAAATCAACAAGAAGAAGTTG 60
frCdx2 GTGAAGATCTGGTTCCAGAATCGCCGGGCCAAAGAGCGCAAAATCAACAAGAAGAAGCTC 60
       ** ** ** ***** ***** *** * ** ** *** * ********   *******

mmCdx2 CAGCAGCAACAGCAGCAGCAGCAACAACAGCCTCCACAGCCGCCGCCACAACCTTCCCAG 117
hsCdx2 CAGCAGCAACAGCAGCAGCAGCCACCACAGCCGCCTCCGCCGCCACCACAGCCTCCCCAG 120
frCdx2 CAGCAGC--CTGCCTCCTCCACGACCACGCCCACGCCTCCCGCCAGCAC--CGGTGCCAG 116
       *******  * **  *  *  * ** **  ** *  *  *****  ***  *    ****

mmCdx2 CCTCAGCCGGGTGCCCTGCGG-AGCGTGCCCGAGCCCTTGAGTCCTGTGACCTCCTTGCA 176
hsCdx2 CCTCAGCCAGGTCCTCTGAGA-AGTGTCCCAGAGCCCTTGAGTCCGGTGTCTTCCCTGCA 179
frCdx2 CCTCCACGGAAACGGTGGCAGCAGCGTCGCCATGGTGACAAG--CAGCAGCGGCAGTAAC 174
        ****  *          *    ** **  *   *      **  * *   *  *  *

mmCdx2 AGGCTCAGTGCCTGGTTCTGTCCCTGGGGTTCTGGGGCCAGCTGGAGGGGTTTTAAACTC 236
hsCdx2 AGCCTCAGTGCCTGGCTCTGTCCCTGGGGTTCTGGGGCCAACTGGGGGGGTGCTAAACCC 239
frCdx2 GGGCT-GGTTTCTCCTTCCTCCCTTC--CTTTGAACATCAA--AGAGGAGTAC------- 222
        * **  **  **   **   ** *    **       **    * ** **
```

```
mmCdx2 CACTGTCACCCAGTGA 252
hsCdx2 CACCGTCACCCAGTGA 255
frCdx2 ----------------
```

Appendix, Section 1C

Cdx4 nucleotide alignment

```
1st exon Cdx4
hsCdx4 ATGTACGGAAGCTGTCTTTTGGAGAAAGAAGCAGGCATGTACCCGGGCACTCTCATGAGC 60
mmCdx4 ATGTATGGAAGCTGCCTTTTAGAAAAAGAAGCGGGCATGTACCCAGGCACTCTCAGGAGC 60
frCdx4 ATGTATG-----------------------------------TTGGATATATTTTGGA- 23
       ***** *                                     **   * *   *

hsCdx4 CCTGGGGGCGACGGCACAGCTGGGACAGGCGGCACAGGGGGCGGTGGGAGTCCGA-TGCC 119
mmCdx4 CCTGGAGGAAGCAGTACAGCCGGAGTGGGCACCTCTGGGGGCAGTGGTAGTCCTC-TGCC 119
frCdx4 --TAAGGAGAGCGGCAT-GTATCACCAGGGACCAGTAAGAAGATCAAGCATCAACCTGCC 80
         *   *    * * *  *        **   *     *           **    ****

hsCdx4 AGCCTCCAATTTCGCTGCGGCACCGGCTTTCTCGCACTATATGGGGTATCCTCATATGCC 179
mmCdx4 TGCCTCCAACTTTACTGCAGCCCCAGTTTACCCACACTACGTGGGTTACCCTCATATGTC 179
frCdx4 CCCCCAGAACTTTGTTTCAACTCCACAGTATCCTGATTTTACCGGATACCATCATGTGCC 140
         **   ** **   * *  * **    *   *  * *     ** ** * **** ** *

hsCdx4 CAGCATGGATCCTCACTGGCCGTCTCTGGGAGTCTGGGGCTCACCCTACAGTCCCCCGCG 239
mmCdx4 CAACATGGATCCTCACGGGCCTTCGCTGGGAGCCTGGAGCTCACCCTACAGTCCGCCCCG 239
frCdx4 GAACATGGATACGCACGCACAGTCTGCGGGGAGTTGGGGGTCTTCGTACGGCGCTCCACG 200
        * ******* * ***   *  **   ***    *** * **  * *** *  * ** **

hsCdx4 AGAAGACTGGAGCGTGTAT--CCTGGG-CCGTCTAGTACAATGGGCACAGTGCCGGTGAA 296
mmCdx4 GGAAGACTGGAGTACATAC--CCTGGG-CCGCCCAGTACAATGGGCACAGTGCCCATGAA 296
frCdx4 AGAGGATTGGGGTGCATACAGCCTGGGACCACCTAATACTAT---TCCCGCCCCTATGAG 257
        ** ** *** *    **   ****** **  * * *** **     * *  **  ***
```

```
hsCdx4 CGACGTGACCTCTAGCCCCGCCGCTTTCTGCTCGACCGACTACAGCAACTTGGGCCCTGT 356
mmCdx4 TGACATGACCT------CCCCAGTTTTCGGATCCCCAGACTACAGCACTCTGGGCCCCAC 350
frCdx4 CAACTCATCTCCCGGACAGGTTCCATACTGCTCACCTGAGTACAGTCATATGCACCCTCC 317
         **    *                * * * **  * ** *****     **  ***

hsCdx4 GGGCGGTGGAACTAGCGGCAGCAGCCTACCAGGCCAGGCTGGCGGGTCGCTTGTCCCGAC 416
mmCdx4 CAGCGGTGCAAGCAACGGCGGTAGCTTGCCAGACGCGGCCAGCGAGTCACTGGTTTCTCT 410
frCdx4 AGGATCTGCGGC--GCTACAGCCGCCT-CCGGAAAACGTCTCTGTTGCGCAACTTTCTCC 374
         *   **       *  * *  ** * ** *     *     *   * *   *  *

hsCdx4 GGACGCAGGCGCCGCCAAGGCCAGTTCCCCCAGCAGGAGCCGCCACAGCCCCTATGCATG 476
mmCdx4 TGACTCCGGCACCTCAGGCGCCACTTCTCCCAGCAGGAGCCGTCACAGCCCCTACGCATG 470
frCdx4 GGACAGAGAAAGACTTTCTTTCCAGTGGATGAATAAAA--CTGCGCAATCCTCTTCCACA 432
        ***   *            *   *     *  *  *  *  * **  **     **

hsCdx4 GATGCGCAAGACGGTGCAGGTGACGG 502
mmCdx4 GATGCGCAAAACTGTGCAGGTGACGG 496
frCdx4 G------------------------- 433

2nd exon Cdx4
hsCdx4 GGAAAACCAGGACAAAAGAAAAGTATCGTGTAGTTTACACTGATCATCAAAGATTGGAGC 60
mmCdx4 GGAAAACCAGGACAAAAGAAAAGTATCGTGTGGTCTACACAGATCATCAACGGCTGGAGC 60
frCdx4 GAAAAACAAGAACGAAGGAGAAGTACAGAGTGGTTTATACAGACCACCAGAGGCTGGAGC 60
       * ***** ** ** ** ** *****  * ** ** ** ** ** ** ** **  *  ******

hsCdx4 TGGAAAAGGAATTCCATTGCAATAGATATATCACCATCCAGAGAAAATCAGAGCTGGCAG 120
mmCdx4 TGGAAAAGGAATTTCACTGCAATAGATACATCACCATCAGGAGGAAGTCAGAGCTGGCAG 120
frCdx4 TGGAGAAAGAGTTTCATTTCAACAGATACATCACCATCAGGAGGAAATCTGAACTGGCTG 120
       **** ** ** ** ** * *** ***** *********  *** ** ** ** ***** *

hsCdx4 TTAACCTGGGCCTTTCCGAGAGACAG 146
mmCdx4 TTAACCTGGGCCTTTCTGAGAGACAG 146
frCdx4 TCAGCCTCGGCCTGTCGGAGAGACAA 146
       * * *** ***** ** ********
```

```
3rd exon Cdx4
hsCdx4 GTGAAAATCTGGTTTCAGAATCGCAGAGCCAAGGAGAGAAAGATGATCAAAAAGAAAATC 60
mmCdx4 GTGAAAATCTGGTTTCAGAATCGCAGAGCCAAGGAGAGGAAGATGATAAAAAAGAAAATC 60
frCdx4 GTGAAAATCTGGTTCCAGAATCGCAGAGCTAAAGAGAGGAAGCTGATCAAGAAGAAGCTG 60
       ************** ************** ** ***** *** **** ** *****  *

hsCdx4 TCCCAGTTTGAGAATAGTGGAGGCTCGGTGCAAAGTGACTCTGACTCCATCAGCCCTGGG 120
mmCdx4 TCCCAGTTTGAGAACACTGGAGGTTCCGTGCAAAGTGACTCTGGCTCCATCAGCCCCGGA 120
frCdx4 GGCCAGTCTGATGGCAGCGGCGGGTCAGTGCACAGTGACCCGGGCTCGGTCAGCCCTCTG 120
         ***** ***    *  ** ** ** ***** ****** * * ***  *******

hsCdx4 GAACTACCTAACACTTTTTTCACCACAC-CATCTGCTGTTCG-TGGATTTCAACCTATTG 178
mmCdx4 GAACTGCCTAACGCTTTCTTCACCACTC-CATCTGCTGTCCG-TGGATTTCAGCCAATCG 178
frCdx4 CCCGTGCCCGGTTCTCTCAGTCCCACGGACGTGCACGGTTCTCTGTACCCTCCCCAGGGG 180
           * **     ** *     ****   * *   * ** *  ** *      **    *

hsCdx4 AGATACAGCAGGTTATAGTCTCCGAATGA 207
mmCdx4 AGATACAGCAGGTCATAGTTTCTGAATGA 207
frCdx4 ATGAACCCCCTCCCATCTATCAGGAAT-- 207
       *   **  *     **       ****
```

Appendix, Section 2A

mCdx1 5.2Kb upstream non-coding sequence

5'-
GGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTTGGGGAAGGCGGG
GAGGTGGGGGGAGCGGTGGAGCAGGGAGGGTGGACGCGTGTGTCTGTTTT
CAATTTCAGGGTTTTTTTTTTTCCTCTCTTTCTTTTCCCCCAACTTTCTTGT
TTTTCTCCTCCTTTCCCTCTGCTCCCTCCCTTCACCGGCTCTTTCAGGATGTT
TANAAATTCCGCTGCCCACGCCAGCCACTGAGTCACTTTTATTANAGTGGC
CTGCCCCACTTGCCTGATGCCTGTCAAATGGAGCCTCCCTTTACCATGGTA
CCTGGGGTGCTAGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGCGCGCG
CGCGCGCGCGTTCTCTGTGCCTCAATTTCTCCATCTGTAAAGGAAGTGCTG
ATAATTTGTGTAGTGACCTCACTGAACATCTGCCCTGTTTGAGGACCTGAG
GACCGGTACTGTGTGCTGGGACTCTGTAGTAAGAGGCCACCTGAGTTGGA
GAGACATTCCTAGCTAAGCAGGTCCCC**AAGGAAAAGAGTGGCTGATGA**<u>**C**</u>
ACACTAGAGAAAGACCCAGGGCTGACACTGCTACTGTTGAGTCCTGACCG
TTCACCTGGGCCTGGAGCTTTCTCCCTTGCTACCTTAAACTTTGGGTCTGA
CTCCTGTTCATCTCCACACACAGATGAGTACAGATCAAAGGCGTGTCTGTC
TAAATACTCACTGCAGCTGAAAAACCACTAAGAAGCTGAGAATGGTGGTG
GATGCCTGCCATCCCAGATAGACACAGCAGAAAATGGCTCTCGGCTATCC
AGGGAGACCCTGTGTAAAAACAGTATATTCTTAGATTCCAGAAACAGCAA
ATCTTACAGGAAGATCATTCCAGGAACTTAAGGATGATGGGACTGAGTCC
CCCTCTCAAGCTGTGGTACCGTCCTAGAAGCCAGTGGAACCTCACAGAAC
CAGACGTGCGATCACCATGTGCCAGACCACAGGATAATGGGAGGCAGCCT
GACTTATGGACTCTTTCTTCTTGGTCCCTGCTCTAGATGTGTCCCCCTCCAT
CTCTTGTTAATACCACAGACAGG**ATGTAGGGCAATAGTT**CCCATCTCTTC
CCA<u>**G**</u>TGTGACTGCACTGAGATATAAACACTCTTCCGCACTGTCATATGTCT
GACTGGACTGTCCAGGCAGCAGTGACTGAGTCTGGTGTCAGGTTCCTGAA
ACTCTCCCAATAAGGATAATTCCAGTCTTGTTGTTTTGCTCTGTTTTCTAGA
AGAAACAAAGGCCCAGAGGTAAGTGGCACAGAGAGAGAGAGACCATGAC
TACCTGAGGTCACCCAGTCAGTACATTGCCCCATGTGCTCAGGCAGACCC
CCTTTAGCATCTTAAGGCTCCCGAGGAAACAGGAAGAGATTCTTTGCCCT
GTTTTATAGCAGGGCAACTGAGGTACAAGTTCCGTGAGTTTCCTGAGCTCG
CGCAGAAACAGCAAAGGTGGGACTAAGTTGTCTGGAACCACAGGAGCAG
AGACAGGCGCCCTGCCCCACCCTCCTCCTAGCCTACCCAGCAGTGATCCA
GCGAGGTATCTCCCATTGTGTTCCTGCCTACCCCTGCCCCTCTCCACAGAG
GTCCTGGCTGG**TTCTCTTCCACTTTC**GCCTTCCTAGATCCCTCCCCCA<u>**T**</u>TT
TNCGTGCCCTAGCTCTCAAAGCCACACTGGGTGTCCTTTCAAAAGGACTGT
TAGACTCTAGTGAACGCATACAGTGTCACTTCTGGGGAAAGTTGGGTTTC
AGGGTTTGGAGACGGGGTATCCCAGATCAGGAACAGCTAGGTTCTGCCCT
CTTTACACCCTATCATCTGTGTCTTTGTGCCCCGCGCCCCTCCTCTTCTCCA
GTCCCACAAAGGTCTCAGGTCAGGTACATCTGAGATTCCATTATTTATCTT
GCTTTCGTTTCTTGGTCTCATCTTCTTAGCAATTACTTAGCAAAGGACTGG
GCGGGGCTGTGCGGAAAAGGGGATTGAAACCGGCCCAGACCATGTGAGG
TAATGAATCTCTAGGTGGAGTGCCTGACCCAGCG**GTCATAGAGGAGC**AT
GGATGATGCCAAAGGTCACAGGGTCACTGGGCAAGCCTA<u>**C**</u>ACAGGACTTC
TTGTACAAAGGCAAACTCCCTGGGTTTGAGCCCCAAACCACAAAACAGAA
AGAGGTGCACGGGGCAGGAAACCGTGTCCCCAGAATTTCACTGGCTGTGC

TTCTCTGAACCCTCAGATCAATTGCAGTGAGCAGGATAGCTTGTAAGAAG
CAAGCTATAGTCCATAGTCCCAAAAGACGTGAGAACACCTGGCCCTGGAG
TTTCACAATCACTGTATGGGTCCTTATCCTGTCTGAGCCTCAGTTTCTCAGC
ACAAGGACCTTGGGGATGACCTAGGGAAAGGAGACTGGCAGGGCTTGGG
GCAGGGGACAGCTGGGGGCCA**AAGACACAGCGTAGA**GGGAAGGAGTGG
CGGTGGCCGATCCTCGAGAGATTGGGGGGGGGGGCTG**C**ATTTCCGGCCAG
CCCAGCCATCCGCTGGCCCATCTGGAAGCACTTAGTTCTACCACCATTAAC
CCTTGGTGGGCACAGGAGCTGTTTCTCCTCCCTGTGATGGGATGCAAGCG
ATACAAAACTACACAGAGTGCCACAGGGCACTCTCGAAGGCCTGGACCCA
GANACTAAAGTTGAGTTCTGACTCCAGCCTCTCTGTTTGAAAGCTGGAAA
GACCTTGTTGAATGTCCCTGAGCTGCCCAGCCTGCTTTCTGCTCTAG**ACAG**
CAGGTAAGTCAGGGCCGGGTCTTGCACCCGCAG**C**TTCTTAGCAGTCTGTG
GGTGAACGAGTATGTAGCATCACGCTAGGGTGGCCAGCATGGTCCCAGCC
CCACTGTGGACAGAACGAATCAAGCACACTCATTCCACGCCCAGATCCTG
GGAGCCGGTTTACCTATGGTCAAACACTGAGGGAAAAGCTGCTGAGAGCG
GGTGGTGGGTAGAGTTCAGAAGGGGCTCCAGGCCTTGGGGCGTCACGCTT
TTTTCTCACCTACGCAGATTGGCGGTCTACTTCATGCATGAGAATCTGAGC
GCTAAGCCTGAGGCTCCGAACTCAGTTGCGTCTTAGCTGTGTGACTCTCAA
AATGCCCAGAGCCTTGGGGTTTTTAAAATAAAGGCTTGTGTGCCCACCTCA
CATGCTGTGAGGGACATGAGTACAAGATTCATACAGGTTTGGGGCGAGCC
TAGAGGCAGAAGGAGAGGAGAACTACAAAGGATAACCAGTCCTCCTGGG
GCTGCTGTCACAG**GAGTCCATAATCAAGG**GCCGGAAGGACCCCCCTTCA
CCC**A**CTCCCTTTATAGCCTACGTAAGGGGTCTNGACAGAGTGGCTCATGTG
AGGCTGTATGATCCTGCAGTGTGATCACAGATACCATGGCAACAAAGGCT
TGGCCCCNNGTTTTATACCCACGTTCTGCCCTAGAGGAGGGGTCCAGGTTG
CGCATAATCCCTGATACCTCCCTCACTGTCAGAGCAGGAAGCAGGGCTCT
GTGGAGTGTGGGCTGAGGGGCCCGGAAGGGGATCCCCATGGGAGATGTC
CCTGTGGATCAGCTCAGCACCAGCAGCTACCACAGCTGGCTCTCCCACGC
TCCTGAAGCTGCCCTGGAGGCTCGCAGAGCAGGCATGGGGTGTGAGGAAG
CAGCTCACGTCCCAGACCTTAGCACAGTCAGCTTGGGTGTCCGGAAGGAC
GTCCTCTAGGGTCGTCTACTGGACAAGCTCAGGAAATGGAAGGCTTTGCA
GGGAGGTGCCTGAGCGGTAGAGGATGATTTAGCCCAN**CATCCTTCCCTA**
CTACTCACTGTGACAGATAGGTAGACCCTA**C**CTATAGCTGTTCTACCTGTA
GCTGGAAATCAAGTCTTAGGGTAGGAGTCACAGCACCGTTGACACTTGGA
GCTCGTGTCCGAGGGAGGCAGCAGGGACATGAGTCATCTTCAACAGAGAT
TCTTGGAAACCCTCTTCCCGGGCCAGGCTCCACTAAAGCAGGACATTTGA
AGTTCAAGACATGTCCATTCATACTTACCCCAGCCTCCCAACCACCCTATC
CTATGCTACCTCAGTGCCCCCGGAACACAGGGCACTGTGGCACATCCTCCT
ACCATATACAGAAGTCTCAAAGAAATACAGCAGTAGCACCCTTGAAGACG
ATATTGTTTCTCCCTCACGAGAACACTGGCTCTTAAGACCCAGAAGTGGGT
ACAGAACCTGACTGAAGGCACATAATATGTCTTCCGTGGGGCCTCTCACC
CCTAATTTCAGGCTGTCCTGTTTCTCCCATTGTGCATCTGCATGAAGGGAG
TTAGTTCGCTGAAC**ACCTACTATGTGACCT**GCCTTCCCACGCCCCTTCCC
AGAGCTNACAAGCA**A**AcTTCAAGGAAGCAACCGATGAATAACACCAATCT
GTGGTTGGTACAATGAAAGATACAGAAATCACAGAGTTGGGCAGATGGG
GGTGGGGGGTCATGCTCAGGCCTGTAGAAGCTCCTTTGGGGAAGACGTAT
CTCCGCTCAGACCTGAGGTTGAAAGGCCATCCATGAAACAAGAATGGGAA
CAGTGTTGAAAGCAAAGGGATCAGCCCCTGCAAATTCTTTGAGGTTCACA
AGACCTTGAAGTGGTAGATCAGAAGAGAGGAGAGGGATTGCCAGGATGG
GGAAGGCAGCTGGGGGCCAGCGCAAAGACTTGAGGGCTCTTGGGAGCCC

AGCTTTTATTCTGAGGTCAGCAAGCCGCTGCAgCTCAACATGCATGAGACC
AGGTAGCTGAGGTATAGTGTGGGGGTGTAAGGCTGGGCTGGAgTCTTGGG
GAGCTTGTGGGCAACTGTGAATCANAGTTGGAACCTCTAGCTCTCTCTGA
ACTTTGACCCCTGATCTTCGTGCCTTCTTTCCTCCCTGCCTGTCTGTTTGCA
CATTGTTGGGCACAAGGACATTCTGTGCTTGTTCTNAACCCCTTATTCTGG
TGTCNTCTTTGATCTCTGGTGTCCTGGATTAAGGA-3'

Figure. *mCdx1* 5' upstream regulatory region. The 5.2 Kb fragment was sequenced from the distal upstream sequence of the *Cdx1*gene. The bold characters show the primer position in the sequence.

Appendix, Section 2B

mCdx1 LacZ reporter constructs

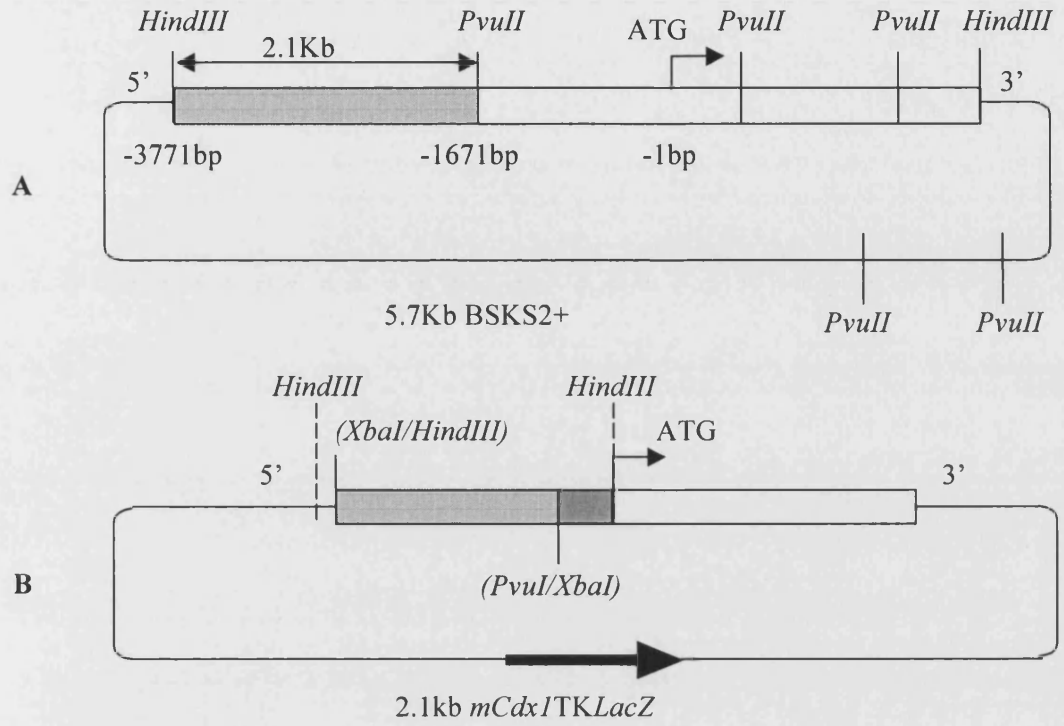


Figure. Plasmid map of 2.1kb *mCdx1*TK*LacZ*. The parent plasmid is shown (A). Digestion with *HindIII/ PvuII* releases the 2.1kb fragment. The dotted square indicates the 2.1kb *Cdx1* fragment. (B) The 2.1kb *Cdx1* fragment cloned into the *XbaI* site of the pTK*LacZ* vector. The grey square represents the TK promoter and the open square shows the *LacZ* gene. The ampicillin resistance gene is indicated by the black arrow.

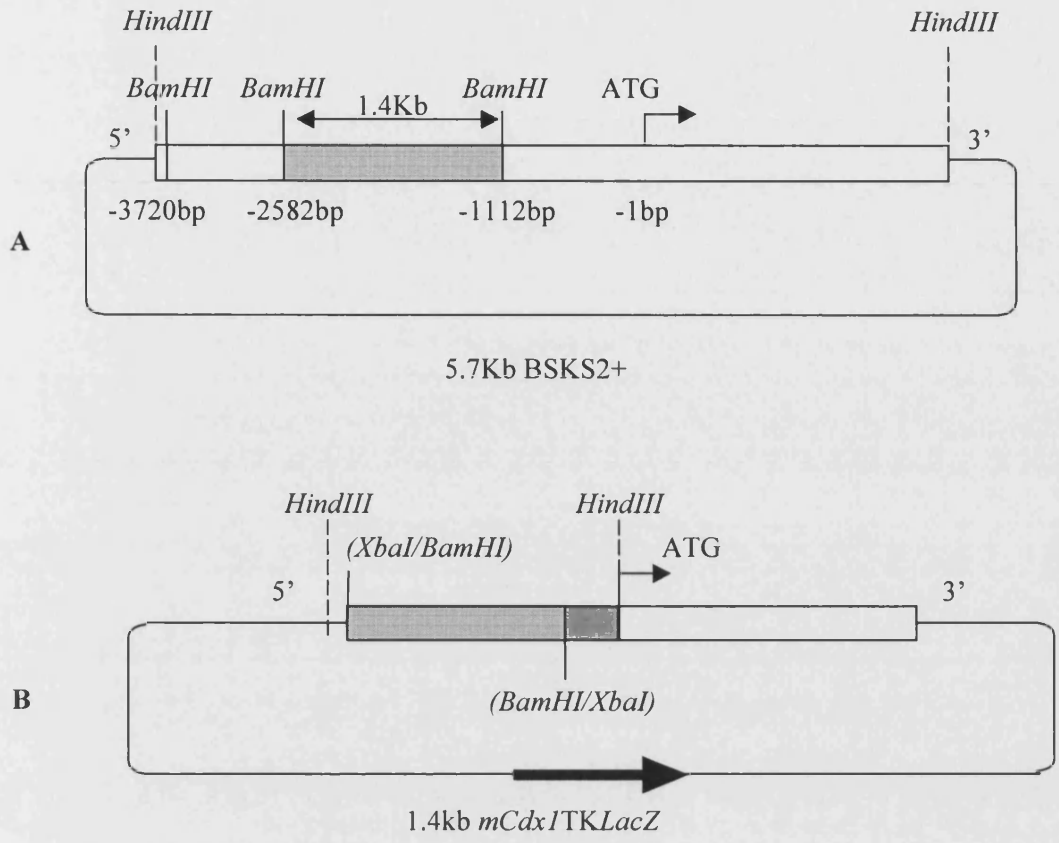


Figure. Plasmid map of the 1.4kb *mCdx1*TK*LacZ*. The parent plasmid is shown (A). Digestion with *BamHI* releases the 1.4kb fragment. The dotted square indicates the 1.4kb *Cdx1* fragment. (B) The 1.4kb *Cdx1* fragment cloned into the *XbaI* site of the pTK*LacZ* vector. The grey square represents the TK promoter and the open square shows the *LacZ* gene. The ampicillin resistance gene is indicated by the black arrow.

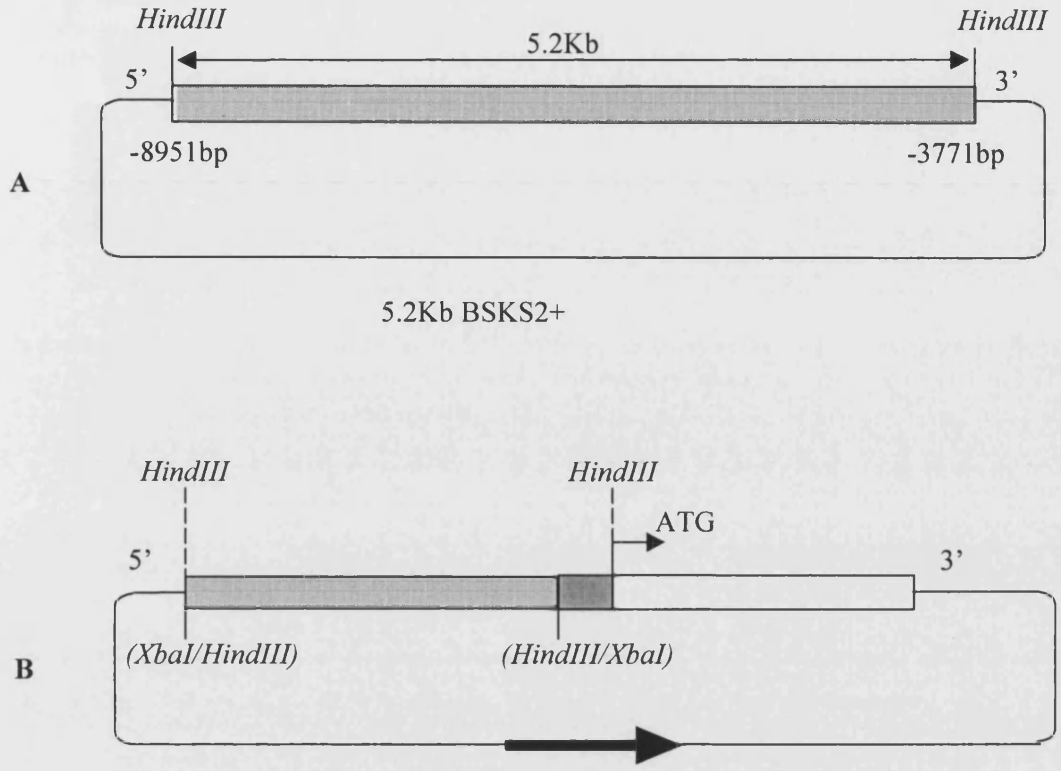


Figure. Plasmid map of 5.2kb *mCdx1*TK*LacZ*. The parent plasmid is shown (A). Digestion with *HindIII* releases the 5.2kb fragment. The dotted square indicates the 5.2kb *Cdx1* fragment. (B) The 5.2kb *Cdx1* fragment cloned into the *XbaI* site of the pTK*LacZ* vector. The grey square represents the TK promoter and the open square shows the *LacZ* gene. The ampicillin resistance gene is indicated by the black arrow.

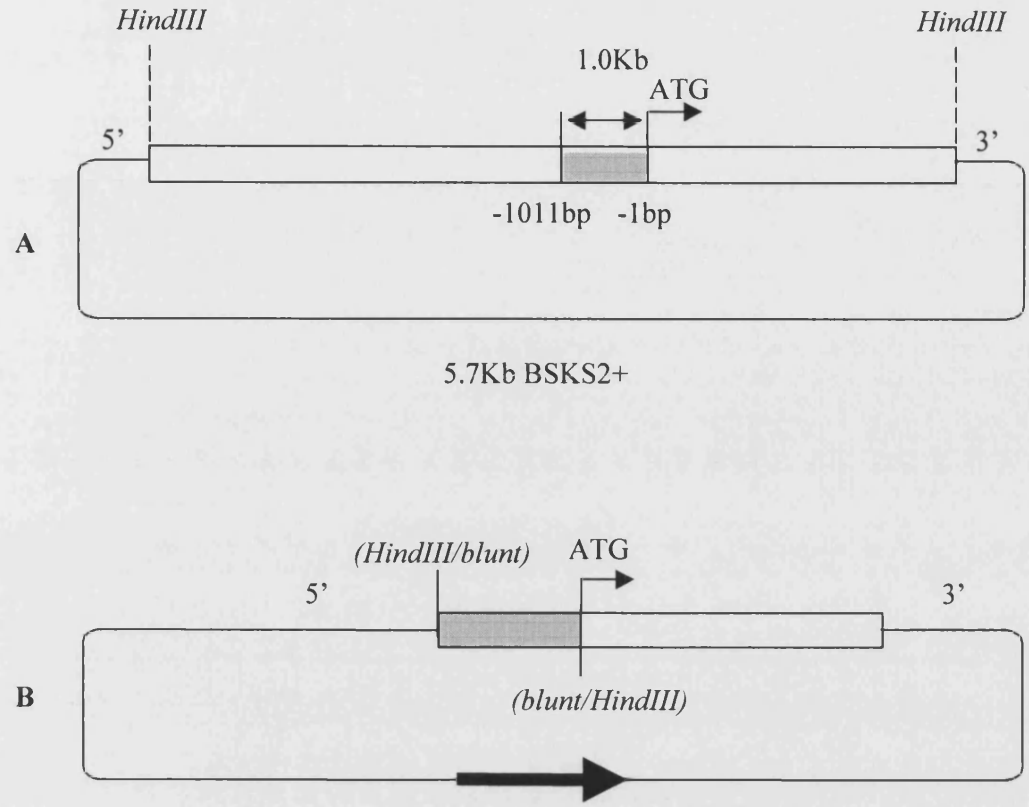


Figure. Plasmid map of 1.0kb *mCdx1*TK*LacZ*The parent plasmid is shown (A). The dotted square indicates the region were the 1.0Kb PCR product was amplified. (B) The 1.0kb *Cdx1* PCR fragment was cloned into the *HindIII* site of the pTK*LacZ* vector. The open square shows the *LacZ* gene. The ampicillin resistance gene is indicated by the black arrow.

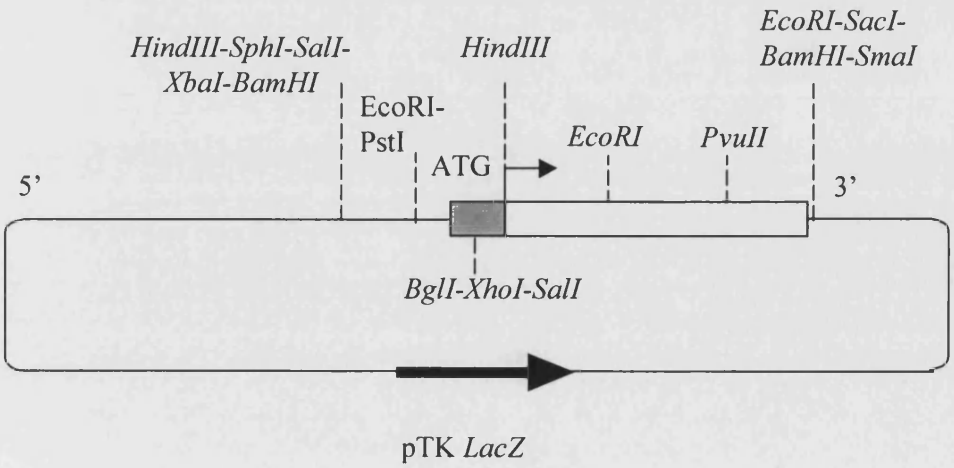
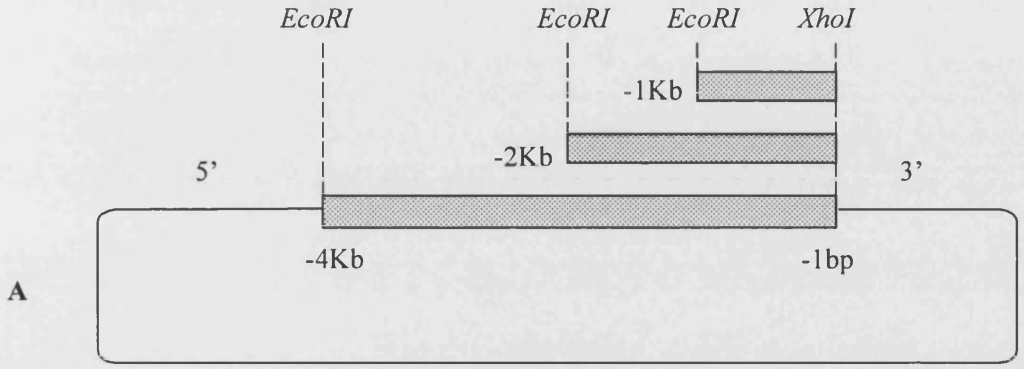


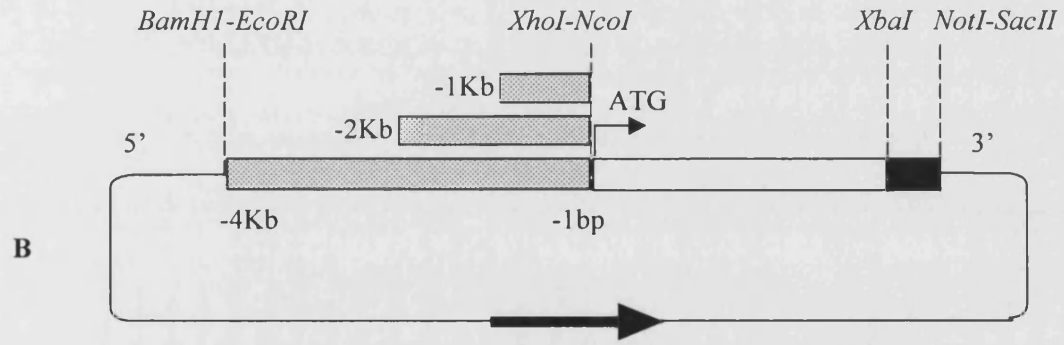
Figure. Plasmid map of pTK*LacZ*The parent plasmid is shown (A). The grey square represents the TK promoter and the open square shows the *LacZ* gene. The ampicillin resistance gene is indicated by the black arrow.

Appendix, Section 2C

frCdx1 GFP reporter constructs



frCdx1 PCR fragments- BSKS2+



frCdx1 deletion constructs- pCS2:GFP^{-CMV}

Figure. Plasmid map of *frCdx1* deletion constructs- pCS2:GFP^{-CMV}. (A). Shows the cloning of the frCdx1 PCR products into BSKS2+ using the *EcoRI/ XhoI* sites. The dotted square indicates the 4.0, 2.0 and 1.0 Kb fr*Cdx1* PCR products. (B) Cloning of the fr*Cdx1* PCR products into the pCS2:GFP$_{-CMV}$ vector. The open square shows the *GFP* gene and the black square indicates the SV40polyA. The ampicillin resistance gene is indicated by the black arrow.

Appendix, Section 3A

Apc1 nucleotide alignment

```
E1hs ATGGCTGCAGCTTCATATGATCAGTTGTTAAAGCAAGTTGAGGCACTGAAGATGGAGAAC 60
E1mm ATGGCTGCAGCTTCATATGATCAGTTGTTAAAGCAAGTTGAGGCACTGAAGATGGAGAAC 60
E1fr ATGGCGGCAGCGTCATACGACCAGCTGCTAAGGCAGGTGGAGGTGTTGAAGATGGAGAAC 60
     ***** ***** ***** ** *** ** *** *** ** ****   **************

E1hs TCAAATCTTCGACAAGAGCTAGAAGATAATTCCAATCATCTTACAAAACTGGAAACTGAG 120
E1mm TCAAATCTTCGACAAGAGCTAGAAGATAATTCCAATCATCTTACAAAACTGGAAACTGAG 120
E1fr TCTAACCTGCGACAGGAGCTGCAGGACAACTCCAACCATTTGACCAAACTGGAGACGGAA 120
     ** ** ** ***** *****  * ** ** ***** *** * ** ******** ** **

E1hs GCATCTAATATGAAG 135
E1mm GCATCTAATATGAAG 135
E1fr GCCTCCAACATGAAG 135
     ** ** ** ******

E2hs GAAGTACTTAAACAACTACAAGGAAGTATTGAAGATGAAGCTATGGCTTCTTCTGGA--- 57
E2mm GAAGTACTTAAGCAGCTACAGGGAAGTATTGAAGATGAGACTATGACTT---CTGGA--- 54
E2fr GAAGTGCTGAAGCAGCTGCAGGGCACCATAGAAGAAGAGTCTGGAGAGGCATCTGGATCT 60
     ***** ** ** ** ** ** ** *  ** ***** **  **          *****

E2hs CAGATTGATTTATTAGAGCGTCTTAAAG 85
E2mm CAGATTGACTTACTAGAGCGTCTTAAAG 82
E2fr CAGCTTGAGCTCATCGAAAGACTGAAGG 88
     *** ****  *  * **  * ** ** *

E3hs AGCTTAACTTAGATAGCAGTAATTTCCCTGGAGTAAAACTGCGGTCAAAAATGTCCCTCC 60
E3mm AATTTAACTTAGATA---GTAATTTCCCCGGAGTGAAACTACGCTCAAAAATGTCCCTTC 57
E3fr AAATGAGCCTCGA-ATCAGCAGGCTTCA-AACACAGGACT-CGGCCTCCGATGCCCACCT 57
     *  * * * * ** *   * *   * *          *** **  *    *** **
```

```
E3hs GTTCTTATGGAAGCCGGGAAGGATCTGTATCAAGCCGTTCTGGAGAGTGCAGTCCTGTTC 120
E3mm GCTCCTACGGAAGTCGGGAAGGATCTGTATCCAGCCGTTCAGGAGAATGCAGTCCTGTCC 117
E3fr CCTCTCCCAGTGCCTCTG---GTTCTGGAGCTCCTGGTGCAGCAGGTGGAGGCCCCCAGG 114
       **     *       *   * **** * *     ** * * **   *  * **

E3hs CTATGGGTTCATTTCCAAGAAGAGGGTTTGTAAATGGAAGCAGAGAAAGTACTG---GAT 177
E3mm CCATGGGGTCATTCCCAAGAAGAACATTTGTAAATGGAAGCAGAGAGAGTACTG---GGT 174
E3fr CAAGCGCCGCCTTCGGCAGGAGAGGGATGCCGACTGTGGGCAGAGAGAGCCACGACCGCT 174
     * *  *   * **    ** ***    *    * **   ******* **    *   * *

E3hs ATTTAGAAGAACTTGAGAAAGAGAG 202
E3mm ATCTAGAAGAGCTTGAAAAAGAAAG 199
E3fr GCCTGGAGGAGCTGGAGAAGGAGAG 199
       * ** ** ** ** ** ** **

E4hs GTCATTGCTTCTTGCTGATCTTGACAAAGAAGAAAAGGAAAAAGACTGGTATTACGCTCA 60
E4mm ATCATTACTCCTTGCTGATCTTGACAAAGAAGAGAAGGAAAAGGACTGGTATTATGCTCA 60
E4fr GTCTCTCCTGTTGGCTGAACTAGAGAAGGAAGAGAAGGAGAAGGACTGGTATTACGCTCA 60
      **  * **  * ***** ** ** ** ***** ***** ** ********** *****

E4hs ACTTCAGAATCTCACTAAAAGAATAGATAGTCTTCCTTTAACTGAAAAT 109
E4mm ACTTCAGAACCTCACAAAAAGAATAGATAGCCTGCCTTTAACTGAAAAT 109
E4fr GCTGCAGAATCTCACCAAGAGGATCGACAGCCTGCCGCTCACTGAAAAT 109
      ** ***** ***** ** ** ** ** ** ** **  * *********

E5hs TTTTCCTTACAAACAGATATGACCAGAAGGCAATTGGAATATGAAGCAAGGCAAATCAGA 60
E5mm TTTTCCTTACAGACAGACATGACAAGACGGCAGCTGGAGTATGAAGCAAGGCAGATCAGG 60
E5fr TTCACGCTCCAAACAGACAGGAGTCGTCTACAGCTGGAGTTTGAGGCTCGACAGATACGT 60
     **  *  * ** ***** * **   *    **  **** * *** **  * ** **  *

E5hs GTTGCGATGGAAGAACAACTAGGTACCTGCCAGGATATGGAAAAACGAGCACAG 114
E5mm GCTGCAATGGAGGAGCAGCTTGGCACCTGCCAGGACATGGAGAAGCGTGCACAG 114
E5fr TCAGCAATGGAAGAACAGTTGGGCTCCTGTCAGGAAATGGAGAGGAGAGCTCAG 114
       ** ***** ** **  * **  **** ***** ***** *   * ** ***

E6hs CGAAGAATAGCCAGAATTCAGCAAATCGAAAAGGACATACTTCGTATACGACAGCTTTTA 60
```

```
E6mm CGAAGAATAGCCAGGATCCAGCAAATAGAAAAGGACATACTGCGCGTGCGCCAGCTTTTA 60
E6fr GCCCGCGTGTCCCGGATTCAGCAGATCGAGAAAGACATCCTGAGACTTGGAGCTCACTTG 60
         *  *  ** * ** ***** ** ** ** ***** **  *  *  *    *  **

E6hs CAGTCCCAAGCAACAGAAGCAGAG 84
E6mm CAGTCCCAGGCGGCGGAAGCGGAG 84
E6fr CAG---------------------- 63
     ***

E7hs AGGTCATCTCAGAACAAGCATGAAACCGGCTCACATGATGCTGAGCGGCAGAATGAAGGT 60
E7mm AGGTCATCTCAGAGCAGGCATGATGCTGCCTCCCATGAAGCTGGCCGGCAGCACGAAGGC 60
E7fr ----------------------------------------------------GAGTGAAGTT 10
                                                          *   ****

E7hs CAAGGAGTGGGAGAAATCAACATGGCAACTTCTGGTAATGGTCAG 105
E7mm CACGGAGTGGCAGAAAGCAACACCGCAGCCTCCAGTAGTGGTCAG 105
E7fr CAAGCGCTGGGTGACAGCAGTGGATTGGCTGCAGCTCAG------ 49
     ** *   ***  ** * **         *  *   *

E8hs GGTTCAACTACACGAATGGACCATGAAACAGCCAGTGTTTTGAGTTCTAGTAGCACACAC 60
E8mm AGTCCAGCTACACGTGTGGATCACGAAACAGCCAGTGTTTTGAGTTCTAGCGGCACGCAC 60
E8fr ACCGCTAGCAGCCGATTGGACCACGAACCAACCAGTGAAGCAAGTT------------AC 48
         *    *  **  **** ** *** ** ******     ****            **

E8hs TCTGCACCTCGAAGGCTGACAAGTCATCTGGGAACCAAG 99
E8mm TCTGCTCCTCGAAGGTTGACAAGTCATCTGGGGACAA-- 97
E8fr TCTGTGCCTCGCCGAATTACAAACCACCTGGGAACAAAA 87
     ****  *****  *  * ****  ** ***** ** *

E9hs --GTGGAAATGGTGTATTCATTGTTGTCAATGCTTGGTACTCATGATAAGGATGATATGT 58
E9mm AGGTGGAAATGGTGTATTCCTTGTTGTCAATGCTTGGTACTCATGATAAGGACGATATGT 60
E9fr --GTGGAGATGGTGTACAGCCTTCTGTCCATGTTGGGGACTCATGATAAAGATGACATGT 58
       ***** ********     *  **** *** * ** ********** ** ** ****

E9hs CGCGAACTTTGCTAGCTATGTCTAGCTCCCAAGACAGCTGTATATCCATGCGACAGTCTG 118
E9mm CACGAACTTTGCTAGCTATGTCCAGCTCCCAAGACAGCTGTATATCCATGCGGCAGTCTG 120
```

```
E9fr CCCGGACTCTCCTCGCCATGTCAAGCTCACAGGACTCCTGCATTGCTATGCGTCAATCAG 118
     * ** *** * ** ** ***** ***** ** ***  *** **  * ***** ** ** *

E9hs GATGTCTTCCTCTCCTCATCCAGCTTTTACATGGCAATGACAAAGACTCTGTATTGTTGG 178
E9mm GATGTCTTCCTCTCCTCATCCAGCTTTTACATGGCAATGACAAAGACTCTGTATTGTTGG 180
E9fr GTTGTCTTCCACTACTCATCCAGTTGCTGCATGGCAATGACAAAGACTCTATTTTGCTAG 178
     * ******** ** ********* *  * ******************** * *** * *

E9hs GAAATTCCCGGGGCAGTAAAGAGGCTCGGGCCAGGGCCAGTGCAGCACTCCACAACATCA 238
E9mm GAAATTCCCGGGGCAGTAAAGAGGCTCGGGCCAGGGCCAGTGCAGCACTCCACAACATCA 240
E9fr GTAACTCTCGAGGAAGCAAGGAAGCCCGCGCGAGGGCATCCGCGGCGTTGCACAACATTG 238
     * ** ** ** ** ** ** ** ** ** ** *****    ** **  * ********

E9hs TTCACTCACAGCCTGATGACAAGAGAGGCAGGCGTGAAATCCGAGTCCTTCATCTTTTGG 298
E9mm TTCACTCACAGCCTGATGACAAGAGAGGCAGGCGTGAAATCCGAGTCCTTCATCTTTTGG 300
E9fr TTCACAGCCAGCCAGATGATAAAAGAGGAAGACGAGAGATCAGAGTGCTTCACCTGCTGG 298
     *****   ***** ***** ** ***** ** ** ** *** **** ***** **  ***

E9hs AACAGATACGCGCTTACTGTGAAACCTGTTGGGAGTGGCAGGAAGCTCATGAACCAGGCA 358
E9mm AACAGATACGAGCTTACTGTGAAACCTGTTGGGAGTGGCAGGAAGCCCACGAACAAGGCA 360
E9fr AGCAGGAGCGACATTACTGCGAGGCCTGTTGGAGTTGGCAGGAGAACCACGAGAGGGGCA 358
     * ***   **   ****** **  ********   ********    ** **    ****

E9hs TGGACCAGGACAAAAATCCAA 379
E9mm TGGACCAGGACAAAAACCCAA 381
E9fr TCGACCAGGAGGACAACCCA- 378
     * ********  * ** ***

E10hs TGCCAGCTCCTGTTGAACATCAGATCTGTCCTGCTGTGTGTGTTCTAATGAAACTTTCAT 60
E10mm TGCCAGCTCCTGTTGAGCATCAGATCTGTCCTGCTGTGTGTGTTCTAATGAAGCTTTCAT 60
E10fr TGCCTTCTCCAGTGGAGCACCAAATCTGTCCAGCCGTCTGTGTCCTGATGAAGCTTTCCT 60
      ****  **** ** ** ** ** ******** ** ** ***** ** ***** ***** *

E10hs TTGATGAAGAGCATAGACATGCAATGAATGAACTAG 96
E10mm TTGATGAAGAGCATAGGCATGCAATGAATGAACTTG 96
E10fr TTGATGAAGAGCATCGACATGCCATGAATGAACTTG 96
```

```
      ************** * ***** *********** *

E11hs GGGGACTACAGGCCATTGCAGAATTATTGCAAGTGGACTGTGAAATGTATGGGCTTACTA 60
E11mm GGGGACTGCAGGCCATTGCAGAGTTATTGCAGGTGGACTGTGAGATGTATGGGCTTACTA 60
E11fr GTGGGCTGCAGGCGGTGGCAGAGCTGCTGCAGGTGGACTGTGAGATGTTTGGTCTCACCA 60
      * ** ** *****  * *****  *  **** ********** **** *** ** ** *

E11hs ATGACCACTACAGTATTACACTAAGACGATATGCTGGAATGGCTTTGACAAACTTGACTT 120
E11mm ATGACCACTACAGTGTTACTTTAAGACGGTATGCTGGAATGGCTTTGACAAACTTGACCT 120
E11fr GCGATCATTACAGCATCACACTGAGGAGATATGCTGGCATGGCCCTCACTAACCTCACAT 120
        ** ** *****  * **  * **  * ******** *****  * ** *** * ** *

E11hs TTGGAGATGTAGCCAACAAG 140
E11mm TTGGAGATGTTGCCAACAAG 140
E11fr TTGGAGACGTCGCCAATAAG 140
      ******* ** ***** ***

E12hs GCTACGCTATGCTCTATGAAAGGCTGCATGAGAGCACTTGTGGCCCAACTAAAATCTGAA 60
E12mm GCTACGCTGTGTTCTATGAAAGGCTGCATGAGAGCACTTGTGGCCCAGTTAAAATCTGAG 60
E12fr GCCACGCTGTGCTCCATGAAGGGCTGCATGAGAGCCATGGTGGCCCAGCTCAAGTCGGAC 60
      ** ***** ** ** ***** **************  * ********  * ** ** **

E12hs AGTGAAGACTTACAGCAG 78
E12mm AGTGAAGACTTACAGCAG 78
E12fr AGTGAGGACCTGCAGCAG 78
      ***** *** * ******

E13hs GTTATTGCGAGTGTTTTGAGGAATTTGTCTTGGCGAGCAGATGTAAATAGTAAAAAGACG 60
E13mm GTTATTGCAAGTGTTTTGAGGAATTTGTCTTGGCGAGCAGATGTAAATAGCAAAAAGACG 60
E13fr GTGATAGCGAGTGTTTTAAGGAACCTGTCGTGGCGTGCTGATGTCAACAGTAAGAAGACA 60
      ** ** ** ******** *****  **** ***** ** ***** ** ** ** *****

E13hs TTGCGAGAAGTTGGAAGTGTGAAAGCATTGATGGAATGTGCTTTAGAAGTTAAAAAG 117
E13mm TTGAGAGAAGTTGGAAGTGTGAAAGCATTGATGGAATGTGCTTTGGAAGTTAAAAAG 117
E13fr TTGCGTGAGGTCGGCAGTGTACGAGCACTGACGGGCTGCGCTCTCGTGGTGCAGAAG 117
      *** * ** ** ** *****   **** *** **  ** *** * *  **  * ***
```

```
E14hs GAATCAACCCTCAAAAGCGTATTGAGTGCCTTATGGAATTTGTCAGCACATTGCACTGAG 60
E14mm GAATCAACCCTCAAAAGCGTTTTGAGTGCCTTATGGAACCTGTCTGCACACTGCACTGAG 60
E14fr GAGTCCACGCTGAAGTCGGTGCTGAGCGCACTGTGGAACCTGTCTGCTCACTGTACGGAG 60
      ** ** ** ** **    **  **** **  * *****  **** ** ** ** ** ***

E14hs AATAAAGCTGATATATGTGCTGTAGATGGTGCACTTGCATTTTTGGTTGGCACTCTTACT 120
E14mm AATAAGGCTGACATCTGTGCTGTGGATGGAGCACTGGCATTTCTGGTTGGCACCCTCACT 120
E14fr AACAAAGCGGACATCTGCGCTGTGGAGGGTGCTCTGGCCTTTCTAGTGGGAACGCTGACC 120
      ** ** ** ** ** ** ***** ** ** ** ** ** *** * ** ** ** ** **

E14hs TACCGGAGCCAGACAAACACTTTAGCCATTATTGAAAGTGGAGGTGGGATATTACGGAAT 180
E14mm TACCGGAGCCAGACAAATACTTTAGCCATTATTGAAAGTGGAGGTGGGATATTACGGAAT 180
E14fr CACTGCAGTCACACTAACACACTCGCCATCATCGAGAGCGGCGGTGGGATCTTGCGAAAT 180
       ** * ** ** ** ** **  * ***** ** ** ** ** ******** ** ** ***

E14hs GTGTCCAGCTTGATAGCTACAAATGAGGACCACAG 215
E14mm GTGTCCAGCTTGATAGCTACAAACGAAGACCACAG 215
E14fr GTTTCCAGCCTTATCGCCACCAATGAGGCGCACAG 215
      ** ****** * ** ** ** ** ** *  *****

E15hs GCAAATCCTAAGAGAGAACAACTGTCTACAAACTTTATTACAACACTTAAAATCTCATAG 60
E15mm GCAAATCCTAAGAGAGAACAATTGCCTACAAACTTTATTACAGCACTTGAAATCTCACAG 60
E15fr GCAGACGCTGCGCGAGCAGGGCTGCCTTCCAACACTGCTTCAGCACCTCAAGTCACACAG 60
      *** *  **  * *** *    ** ** * ***  *  * ** *** * ** ** ** **

E15hs TTTGACAATAGTCAGTAATGCATGTGGAACTTTGTGGAATCTCTCAGCAAGAAATCCTAA 120
E15mm CTTGACAATAGTCAGTAATGCATGTGGAACTTTGTGGAATCTCTCAGCAAGAAATCCTAA 120
E15fr TCTGACCATCGTGTCCAACGCCTGTGGAACGCTCTGGAATCTGTCGGCCAGAGATGCTAA 120
       **** ** **    ** ** ********  * ******** ** ** *** ** ****

E15hs AGACCAGGAAGCATTATGGGACATGGGGGCAGTTAGCATGCTCAAGAACCTCATTCATTC 180
E15mm AGACCAGGAAGCCTTGTGGGACATGGGGGCAGTGAGCATGCTCAAGAACCTCATTCATTC 180
E15fr AGACCAGGAGACGTTATGGGAACTAGGGGCTGTTGGCATGTTGCGCAACCTCATTCATTC 180
      *********  * ** *****  * ***** **  ***** *    **************
```

```
E15hs AAAGCACAAAATGATTGCTATGGGAAGTGCTGCAGCTTTAAGGAATCTCATGGCAAATAG 240
E15mm CAAGCACAAAATGATTGCCATGGGAAGTGCAGCAGCTTTAAGGAATCTCATGGCAAACAG 240
E15fr CAGGCACAAGATGATAGCCATGGGTAGTGCTGCTGCCCTGCGTAATCTGATGGCTAACCG 240
       * ****** ***** ** ***** ***** ** **  *  * ***** ***** **  *

E15hs GCCTGCGAAGTACAAGGATGCCAATATTATGTCTCCTGGCTCAAGCTTGCCATCTCTTCA 300
E15mm ACCTGCAAAGTATAAGGATGCCAATATCATGTCTCCCGGCTCAAGTCTGCCATCCCTTCA 300
E15fr GCCAGCACGCTATAAAGATGCTAGTGTGGTGTCCCCGGGTGCTGGCGCCCCGTCGTTACA 300
       ** **    ** ** ***** * * *  **** ** **  *  *    ** **  * **

E15hs TGTTAGGAAACAAAAAGCCCTAGAAGCAGAATTAGATGCTCAGCACTTATCAGAAACTTT 360
E15mm CGTTAGGAAACAGAAAGCTCTAGAAGCTGAGCTAGATGCTCAGCATTTATCAGAAACCTT 360
E15fr TGTCCGCAAACAAAAGGCATTATTTGAAGAGCTAGACGCCCAGCAGCTGTCCGAGACGTT 360
       **  * ***** ** **  **   *  **  **** ** *****  * ** ** ** **

E15hs TGACAATATAGACAATTTAAGTCCCAAGGCATCTCATCGTAGTAAGCAGAGACACAAGCA 420
E15mm CGACAACATTGACAACCTAAGTCCCAAGGCCTCTCACCGGAGTAAGCAGAGACACAAGCA 420
E15fr TGACAACATTGACAACCTAAGCCCCAAAACGGCACACAGGAAGGGGCGGGGCTGTAA--- 417
       ***** ** *****  **** *****  *  * **  * *    ** * *    **

E15hs AAGTCTCTATGGTGATTATGTTTTTGACACCAATCGACATGATGATAATAGGTCAGACAA 480
E15mm GAATCTTTATGGTGACTATGCTTTTGACGCCAATCGACATGATGATAGTAGGTCAGACAA 480
E15fr ----------------------------------------TAGTGCCAGTGGGACAG---- 434
                                              *  **  * * ** ***

E15hs TTTTAATACTGGCAACATGACTGTCCTTTCACCATATTTGAATACTACAGTGTTACCCAG 540
E15mm TTTCAATACTGGAAACATGACTGTTCTTTCACCATATTTAAATACTACGGTATTGCCCAG 540
E15fr --CCAGCAC-GGCA-------CGTCCGTACACCA-------ACACACCAGTGCTTTCGAG 477
         *  ** ** *        ** * * *****       * **  * **  *  * **

E15hs CTCCTCTTCATCAAGAGGAAGCTTAGATAGTTCTCGTTCTGAAAAAGATAGAAGTTTGGA 600
E15mm CTCTTCTTCCTCAAGGGGAAGTTTAGACAGTTCTCGTTCTGAGAAAGACAGAAGTTTGGA 600
E15fr CCC------------AAAGAATGGAGATGGTTC-------------------------A 499
      * *       *    ***  ****       *

E15hs GAGAGAACGCGGAATTGGTCTAGGCAACTACCATCCAGCAACAGAAAATCCAGGAACTTC 660
```

```
E15mm GAGAGAGCGAGGTATTGGCCTCAGTGCTTACCATCCAACAACAGAAAATGCAGGAACCTC 660
E15fr AAGAGGAT------------------------------CAACGAGGAACCTGGGTAC--- 526
        ****     ****    **    ** **

E15hs TTCAAAGCGAGGTTTGCAGATCTCCACCACTGCAGCCCAGATTGCCAAAGTCATGGAAGA 720
E15mm ATCAAAACGAGGTCTGCAGATCACTACCACTGCAGCCCAGATAGCCAAAGTTATGGAAGA 720
E15fr ----------------------------------GCCCGGAC------------------ 534
                                        **** **

E15hs AGTGTCAGCCATTCATACCTCTCAGGAAGACAGAAGTTCTGGGTCTACCACTGAATTACA 780
E15mm AGTATCAGCCATTCATACCTCCCAGGACGACAGAAGTTCTGCTTCTACCACCGAGTTCCA 780
E15fr GGTGTTTTCCACTAGTGTC--------------------------------CGAGCTTC- 561
       ** *   *** *  *  *                                 **  * *

E15hs TTGTGTGACAGATGAGAGAAATGCACTTAGAAGAAGCTCTGCTGCCCATACACATTCAAA 840
E15mm TTGTGTGGCAGACGACAGGAGTGCGGCACGAAGAAGCTCTGCCTCCCACACACACTCAAA 840
E15fr ------------------------------------------------------------


E15hs CACTTACAATTTCACTAAGTCGGAAAATTCAAATAGGACATGTTCTATGCCTTATGCCAA 900
E15mm CACATACAACTTCACTAAGTCGGAAAATTCAAATAGGACATGCTCTATGCCTTATGCCAA 900
E15fr CAGTGACAGCCTCAACAGCGTGACGAGTGCAGATGG---------------TTACGGCAA 606
      **   ***   ***  *    *   * * ** ** *              *** * ***

E15hs ATTAGAATACAAGAGATCTTCAAATGATAGTTTAAATAGTGTCAGTAGTAGTGATGGTTA 960
E15mm AGTGGAATATAAACGATCTTCAAATGACAGTTTAAATAGTGTCACTAGTAGTGATGGATA 960
E15fr CCGAGGCAAAAATAAACCGTCAACGGA--------GCCGTTTTACTCGT----------- 647
         *   * **   * * ****  **          ** * * * **

E15hs TGGTAAAAGAGGTCAAATGAAACCCTCGATTGAATCCTATTCTGAAGATGATGAAAGTAA 1020
E15mm TGGTAAAAGAGGCCAAATGAAACCCTCAGTTGAATCCTATTCTGAAGATGATGAAAGTAA 1020
E15fr -------------CAGACGAGAGC-------GGAGCC------------------AATAA 669
                  ** * ** * *       * * **                  * ***

E15hs GTTTTGCAGTTATGGTCAATACCCAGCCGACCTAGCCCATAAAATACATAGTGCAAATCA 1080
E15mm ATTTTGCAGTTATGGTCAGTATCCAGCTGACCTAGCCCATAAGATACACAGTGCAAATCA 1080
```

```
E15fr GTGCTGTGTTTACAGGAAGTACCCGGCTGACCTAGCACACAAGATCCGCAGTGCCAATCA 729
       *  **   ***  *  * ** ** ** ******** ** ** ** *  ***** *****

E15hs TATGGATGATAATGATGGAG---AACTAGATACACCAATAAATTATAGTCTTAAATATTC 1137
E15mm TATGGATGATAATGATGGAG---AACTGGATACACCAATAAATTACAGTCTTAAATATTC 1137
E15fr CATGGCAGATGATGACGGAGCAGAGCTGGACACGCCTATCAACTACAGTCTGAAGTACTC 789
       ****  *** **** ****   * ** ** ** ** ** ** ** ***** ** ** **

E15hs AGATGAGCAGTTGAACTCTGGAAGGCAAAGTCCTTCACAGAATGAAAGATGGGCAAGACC 1197
E15mm AGATGAGCAGTTGAACTCAGGAAGGCAGAGTCCCTCACAGAATGAAAGGTGGGCAAGACC 1197
E15fr TGACGAACAGTTGAATTCTGGGAGACAGAGTCC---------------------------- 822
       ** ** ******** ** ** ** ** *****

E15hs CAAACACATAATAGAAGATGAAATAAAACAAAGTGAGCAAAGACAATCAAGGAATCAAAG 1257
E15mm AAAGCATGTGATAGAAGATGAAATAAAGCAAAACGAGCAAAGACAAGCAAGAAGCCAGAA 1257
E15fr ------------------------------------------------------------

E15hs TACAACTTATCCTGTTTATACTGAGAGCACTGATGATAAACACCTCAAGTTCCAACCACA 1317
E15mm CACCAGTTATCCTGTCTATTCTGAGAATACCGATGACAAACACCTCAAATTCCAACCACA 1317
E15fr ---AAGTCACCGCGTCGGTATGGACAGTGATGACGAC----------------------- 856
          * * * *  **   *   ** *     ** **

E15hs TTTTGGACAGCAGGAATGTGTTTCTCCATACAGGTCACGGGGAGCCAATGGTTCAGAAAC 1377
E15mm TTTTGGACAACAAGAATGTGTTTCCCCATATAGGTCAAGGGGAACCAGTGGTTCAGAAAC 1377
E15fr -----GACGAGGAGGACG-----------GCAGGCTGAGGAGAAGGAACGA--CGGTAGC 898
           ***     * * *            ***    ** **   *  *   * * * *

E15hs AAATCGAGTGGGTTCTAATCATGGAATTAATCAAAATGTAAGCCAGTCTTTGTGTCAAGA 1437
E15mm AAATCGAATGGGTTCTAGTCATGCAATTAATCAAAATGTAAACCAGTCTCTGTGTCAGGA 1437
E15fr GACTCGG-----------------------------TGTCGTCCGGTAGCCGC------- 922
       * ***                               ***   ** **    *

E15hs AGATGACTATGAAGATGATAAGCCTACCAATTATAGTGAACGTTACTCTGAAGAAGAACA 1497
E15mm AGATGATTATGAAGATGATAAACCTACCAACTACAGTGAACGTTATTCTGAGGAAGAACA 1497
E15fr -----ATTATTTCGGT------CCCACCA-CCACGAT--ACGTTGTCAC----------- 957
```

```
          * ***   * *      ** ****   *   *  *****

E15hs GCATGAAGAAGAAGA---GAGACCAACAAATTATAGCATAAAATATAATGAAGAGAAACG 1554
E15mm ACATGAAGAAGAAGAAGAGAGACCGACAAATTATAGCATAAAATATAATGAAGAGAAACA 1557
E15fr ---TGCAGCAG------------CAGCAAACTATGGCG------GTGATCCAGCAGGA-- 994
        ** ** **            *  **** *** **        * **  **    *

E15hs TCATGTGGATCAGCCTATTGATTATAGTTTAAAATATGCCACAGATATTCCTTCATCACA 1614
E15mm TCATGTGGATCAGCCTATTGATTATAGTTTAAAATATGCCACTGACATTTCTTCCTCACA 1617
E15fr -------GAGCAGCCAATCGACTACAGCCTGAAATATGGCAGTGACGGTGCC------CA 1041
             ** ***** ** ** ** **  * ******* **  **   * *       **

E15hs GAAACAGTCATTTTCATTCTCAAAGAGTTCATCTGGACAAAGCAGTAAAACCGAACATAT 1674
E15mm AAAACCATCATTTTCATTCTCAAAGAATTCATCAGCACAAAGCACTAAACCTGAACATCT 1677
E15fr TAAAC--CCCTGTTCAAGCCCGAAGAGTCCGCC--------GCAT----------CATCT 1081
       ****   * * ****  * * **** * *  *        ***           *** *

E15hs GTCTTCAAGCAGTGAGAATACGTCCACACCTTCATCTAATGCCAAGAGGCAGAATCAGCT 1734
E15mm CTCTCCAAGCAGCGAGAATACAGCTGTACCTCCATCTAATGCCAAAAGGCAGAATCAGCT 1737
E15fr GTT-------------AAGACTCCCACACCGTCTTC------------GTCGGCTAAACT 1116
       *              ** **  *   ***  * **            *  *  * * **

E15hs CCATCCAAGTTCTGCACAGAGTAGAAGTGGTCAGCCTCAAAAGGCTGCCACTTGCAAAGT 1794
E15mm GCGTCCAAGTTCAGCACAAA---GAAATGGCCAGACTCAAAAAGGCACTACTTGCAAAGT 1794
E15fr CCGCCC---CCCTGC---------------CCCTGCTAACAGGGCCG----TGGCAAAA- 1153
       *  **     * **                *   ** * *  *       * *****

E15hs TTCTTCTATTAACCAAGAAACAATACAGACTTATTGTGTAGAAGATACTCCAATATGTTT 1854
E15mm CCCCTCCATCAACCAAGAAACAATACAGACTTACTGCGTAGAAGACACCCCAATATGTTT 1854
E15fr -------GCTAACCAGGAGTCAACGCAGACCTACTGTGTGGAGGACACGCCCATCTGTTT 1206
             ***** **  ***  ***** ** ** ** ** ** ** ** ** *****

E15hs TTCAAGATGTAGTTCATTATCATCTTTGTCATCAGCTGAAGATGAAATAG--GATGTAAT 1912
E15mm TTCAAGGTGCAGTTCATTATCATCACTGTCATCAGCTGACGATGAAATAG--GATGTGAT 1912
E15fr CTCCCGAGGCAGCTCGCTGTCTTCGCTGTCATCAGAGGAGGAAGAAGAGGTTGATGTCAT 1266
       **  *  * ** **  * ** **  *********  ** ** ***   *  ***** **
```

```
E15hs C---------AGACGACACAGGAAGCAGATTCTGCTAATACCCTGCAAATAGCAGAAATAA 1964
E15mm C---------AGACAACACAGGAAGCAGATTCTGCTAATACTCTGCAGACAGCAGAAGTAA 1964
E15fr AGAGAGGAGAGGTGCTAGCGGAGGGGGAAATGGTGAGTATCCAACAGTTCCTGTCAGCGA 1326
              **     *  *** *  **    *  * **  *  **         *   *

E15hs AAGAAAAGATTGGAACTAGGTCAGCTGAAGATCCTGTGAGCGAAGTTCCAGCAGTGTCAC 2024
E15mm AAGAGAATGATGTAACTCGGTCAGCTGAAGATCCTGCAACTGAAGTTCCAGCAGTGTCCC 2024
E15fr GAAGGATGCTCGTGAGCAG--CAGCAGAGGCACCAGAAGGAGG-----CAG-AGAGTCAG 1378
       *   *     *  *   *  **** ** *  ** *     *      *** ** ***

E15hs AGCACCCTAGAACCAAATCCAGCAGACTGCAGGGTTCTAGTTTATCTTCAGAATCAGCCA 2084
E15mm AGAATGCTAGAGCCAAACCCAGCCGACTCCAGGCTTCTGGCTTATCTTCAGAATCAACCA 2084
E15fr ACTGCAGCCGTCACGGCACCGTCTACACGGGGACGGCGAAGTCACCACCATCACCA-CCA 1437
      *        *   *    **  *        *    *    * * *  **  * ** ***

E15hs GGCACAA---AGCTGTTGAATTTTCTTCAGGAGCGAAATCTCCCTCCAAAAGTGGTGCTC 2141
E15mm GGCATAATAAAGCTGTTGAGTTTTCTTCAGGAGCCAAGTCTCCCTCCAAAAGTGGTGCTC 2144
E15fr CGGTCACCACCACCACC--ACCACCACCACG--TTGCATCATCTTC------AGGTGCTA 1487
       *   *      * *           **  *       **  * **       ******

E15hs AGACACCCAAAAGTCCACCTGA---ACACTATGTTCAGGAGACCCCACTCATGTTTAGCA 2198
E15mm AGACACCCAAAAGTCCCCCAGA---ACACTATGTCCAGGAGACTCCGCTCGTATTCAGCA 2201
E15fr GGACTCCCAAGAGTCCTCCAGAGCCACCATATGCCCAGGAGACACCGCTCATGTTCAGCC 1547
       *** ***** ***** ** **   **  ****  ******** ** *** * ** ***

E15hs GATGTACTTCTGTCAGTTCACTTGATAGTTTTGAGAGTCGTTCGATTGCCAGCTCCGTTC 2258
E15mm GGTGTACTTCTGTCAGCTCCCTTGACAGTTTTGAGAGTCGCTCCATTGCCAGCTCTGTTC 2261
E15fr GCTGCACATCCGTCAGTTCCCTGGACAGCTTTTCAACTTCCTCAATTGCCAGTTCTGTGC 1607
      * ** ** ** ***** ** ** ** ** ***   * *   ** ******** ** ** *

E15hs AG---AGTGAACCATGCAGTGGAATGGTAAGTGGCATTATAAGCCCCAGTGATCTTCCAG 2315
E15mm AG---AGTGAGCCATGTAGTGGAATGGTGAGTGGCATCATAAGCCCCAGTGACCTTCCAG 2318
E15fr GCTCCAGCGAGCCGTGTAGTGGCATGCCGAGTGGTGTTGTCAGCCCTAGTGACCTACCCG 1667
          ** ** ** ** ***** ***   *****  *  * ***** ***** ** ** *
```

```
E15hs ATAGCCCTGGACAAACCATGCCACCAAGCAGAAGTAAAACACCTCCACCACCTCCTCAAA 2375
E15mm ATAGTCCTGGGCAGACCATGCCACCAAGCAGAAGCAAAACCCCTCCACCTCCTCCACAGA 2378
E15fr ACAGCCCGGGACAGACCATGCCGCCAAGCCGTGCAAAGACGCCACCGTTGCCACCGACCA 1727
      * ** ** ** ** ******** ****** *    ** ** ** **    ** **    *

E15hs CAGCTCAAACCAAGCGAGAAGTACCTAAAAATAAAGCACCTACTGCTGAAAAGAGAGAGA 2435
E15mm CAGTGCAGGCCAAGAGAGAGGTGCCAAAAAGTAAAGTCCCTGCTGCTGAGAAGAGAGAGA 2438
E15fr CGCATAAAACCAA-----------CAAGCAGGAAAATACC--------AAGAAAAAAGAT 1768
      *     *  ****          * *  *  ***   **        *  * * *

E15hs GTGGACCTAAGCAAGCTGCAGTAAATGCTGCAGTTCAGAGGGTCCAGGTTCTTCCAGATG 2495
E15mm GTGGGCCTAAGCAGACTGCTGTAAATGCTGCCGTGCAGAGGGTGCAGGTCCTTCCAGACG 2498
E15fr GAGGACGAGAGCAG-----------------------------------------------TG 1784
      * ** *   ****                                                 *

E15hs CTGATACTTTATTACATTTTGCCACGGAAAGTACTCCAGATGGATTTTCTTGTTCATCCA 2555
E15mm TGGATACTTTGTTACACTTCGCCACAGAAAGTACTCCAGACGGGTTTTCTTGTTCCTCCA 2558
E15fr CTGATGTTTTGCTGCACTTTGCTACTGAAAGCACGCCACACGGCTTCTCCAGGGCTTCCA 1844
       ***  ***  * ** ** ** ** ***** ** *** * ** ** **  *  * ****

E15hs GCCTGAGTGCTCTGAGCCTCGATGAGCCATTTATACAGAAAGATGTGGAATTAAGAATAA 2615
E15mm GCCTAAGTGCTCTGAGCCTGGATGAGCCATTTATACAGAAAGATGTAGAATTAAGAATCA 2618
E15fr GTCTGAGTGCGCTTAGTGTAGATGAGCCCTTCATC------ACTGCAGAGTTAA------ 1892
      * ** ***** ** **  * ******** ** **         **  ** ****

E15hs TGCCTCCAGTTCAGGAAAATGACAATGGGAATGAAACAGAATCAGAGCAGCCTAAAGAAT 2675
E15mm TGCCTCCAGTTCAGGAAAACGACAATGGGAATGAAACTGAATCAGAACAGCCTGAGGAAT 2678
E15fr ------------AGGAGGAGGA------GGAGGATAAAGGATCCGAGCAGA--------- 1925
                 ****  * **      * * ** *  * *** ** ***

E15hs CAAATGAAAACCAAGAGAAAGAGGCAGAAAAAACTATTGATTCTGAAAAGGACCTATTAG 2735
E15mm CAAATGAAAACCAGGATAAAGAGGTAGAAAAGCCT---GACTCTGAAAAAGACTTATTAG 2735
E15fr ------------GGGTGGAAGAGGCACCAAAAGCA---GTCCTCGATGA------ATCGG 1964
                   *   ****** *  ***  *    *     **  *      **  *

E15hs ATGATTCAGATGATGATGATATTGAAATACTAGAAGAATGTATTATTTCTGCCATGCCAA 2795
```

```
E15mm ATGATTCTGATGACGATGATATTGAAATATTAGAAGAATGTATTATTTCAGCCATGCCAA 2795
E15fr ACGAC------GACGACGACATCGAGATTCTGGAAGCCTGCATCAACATGGCCATGCCT- 2017
      * **       ** ** ** ** ** **  * ****  ** ** *     ********

E15hs CAAAGTCATCACGTAAAGCAAAAAAGCCAGCCCAGACTGCTTCAAAATTACCTCCACCTG 2855
E15mm CAAAGTCATCACGCAAAGCCAAAAAACTAGCCCAGACTGCTTCAAAATTACCTCCACCTG 2855
E15fr --AAATCATCGCGAAAACCAAAGAAACCACAACAAGCAGC---------------------- 2055
        ** ***** ** *** * ** ** * *   **  * **

E15hs TGGCAAGGAAACCAAGTCAGCTGCCTGTGTACAAACTTCTACCATCACAAAACAGGTTGC 2915
E15mm TGGCAAGGAAACCAAGTCAGCTACCTGTGTATAAACTTCTGCCAGCACAGAATAGGCTGC 2915
E15fr -GCCGCGGAAAGCCAGCCAGCTTCCCGTCTACAAGCT-CCGCGTCCA--------GTCGC 2105
       * *  ***** * ** ***** ** ** ** ** ** *  *   **        *  **

E15hs AACCCCAAAAGCATGTTAGTTTTACACCGGGGGATGATATGCCACGGGTGTATTGTGTTG 2975
E15mm AGGCACAAAAACATGTTAGCTTTACACCAGGGGATGATGTGCCCCGGGTGTACTGTGTAG 2975
E15fr AGCCAAGGAAGGACGT--GCCGC-CGCCGGCCGAGGAGGTGCCAAGAGTCTACTGTGTCG 2162
      *  *    **  * **  *     * ** *  ** **  ****  * ** ** ***** *

E15hs AAGGGACACCTATAAACTTTTCCACAGCTACATCTCTAAGTGATCTAACAATCGAATCCC 3035
E15mm AAGGGACACCTATAAACTTTTCCACAGCAACGTCTCTAAGTGATCTGACAATAGAGTCCC 3035
E15fr AGGGAACGCCACTGAATTTCTCCACCGCTACCTCACTGAGTGATCTCACCATCGACTCCC 2222
      * ** ** **  * ** ** ***** ** ** ** ** ******** ** ** ** ****

E15hs CTCCAAATGAGTTAGCTGCTGGAGAAGGAGTTAGAGGAGGGGCACAGTCAGGTGAATTTG 3095
E15mm CTCCAAATGAATTGGCTACTGGAGATGGGGTCAGAGCGGGTATACAGTCAGGTGAATTTG 3095
E15fr CACCCAACGA----------GGAGGCGGCGGCGGTGGCGGCG------------------ 2254
      * ** ** **          ****  ** *   * *  **

E15hs AAAAACGAGATACCATTCCTACAGAAGGCAGAAGTACAGATGAGGCTCAAGGAGGAAAAA 3155
E15mm AAAAACGAGATACCATTCCTACAGAAGGCAGAAGTACAGATGATGCTCAGCGAGGAAAAA 3155
E15fr -----------GCCATCCTGCCCGAGGGCCCGCCCGC--------CTCCACTCAGGAAGA 2295
                ****  *   * ** ***       *        ***      * ** *

E15hs CCTCATCTGTAACCATACCTGAATTGGATGACAATAAAGCAGAGGAAGGTGATATTCTTG 3215
E15mm TCTCATCTATAGTTACACCAGACCTGGATGACAACAAAGCAGAGGAAGGAGATATTCTTG 3215
```

```
E15fr TCGGGCC---GGACATCCC-------GAAGGTGAAAATGC-------GATGACATCCTTG 2338
       *    *       *  **       ** *   * ** **       *  ** ** ****

E15hs CAGAATGCATTAATTCTGCTATGCCCAAAGGGAAAAGTCACAAGCCTTTCCGTGTGAAAA 3275
E15mm CAGAATGTATCAATTCTGCTATGCCCAAAGGAAAAAGCCACAAGCCTTTCCGAGTGAAAA 3275
E15fr CCGAATGTATCAGCGCTGCTATGCCCAAAGCCAAACCCAGGAAGCCAATCAGAGTGGCAG 2398
      * ***** ** *   ***************  ***      *****  ** * ***  *

E15hs AGATAATGGACCAGGTCCAGCAAGCATCTGCGTCTTCTTCTGCACCCAACAAAAATCAGT 3335
E15mm AGATAATGGACCAAGTCCAACAAGCATCCTCGACTTCATCTGGAGCTAACAAAAATCAAG 3335
E15fr TGAACAGCGAGCACGTCCAATCGCCACCC-CTCCCCCCTCCGAAA--AAAAAAAAAAAAG 2455
       **  *  ** ** *****     ** *  *  *  * ** * *   ** *****  *

E15hs TAGATGGTAAGAAAAAGAAACCAACTTCACCAGTAAAACCTATACCACAAAATACTGAAT 3395
E15mm TAGACACTAAGAAAAAGAAGCCTACTTCACCAGTAAAGCCCATGCCACAAAATACTGAAT 3395
E15fr CAGCT--CCAAAAGAAAAAGCCCACCTCCCCTGTGAAGCCAATGCCACAGAG-------- 2505
       **      * ** ** ** ** ** ** ** ** ** ** ** ** ***** *

E15hs ATAGGACACGTGTAAGAAAAAATGCAGACTCAAAAAATAATTTAAATGCTGAGAGAGTTT 3455
E15mm ATAGAACGCGTGTGAGAAAGAATACAGACTCAAAAGTTAATGTAAATACTGAAGAAACTT 3455
E15fr ------------------------------------GGTGGTGTACACCATGA-------- 2522
                                           *  * ** *   ***

E15hs TCTCAGACAACAAAGATTCAAAGAAACAGAATTTGAAAAATAATTCCAAGGTCTTCAATG 3515
E15mm TCTCAGACAACAAAGACTCAAAGAAACCAAGCTTACAAACCAATGCCAAGGCCTTCAATG 3515
E15fr -CCACGACTACA---------------------------------------------- 2533
       *   *** ***

E15hs ATAAGCTCCCAAATAATGAAGATAGAGTCAGAGGAAGTTTTGCTTTTGATTCACCTCATC 3575
E15mm AAAAGCTACCTAACAATGAAGACAGAGTGCGGGGGAGCTTCGCCTTGGACTCACCGCATC 3575
E15fr ----GCTGCCAAA--------ACAAAGCCAGGG-----TTTGCCTTTGATTCACCTCGCC 2576
          *** ** **        * * **   * *     ** ** ** ** ***** *  *

E15hs ATTACACGCCTATTGAAGGAACTCCTTACTGTTTTTCACGAAATGATTCTTTGAGTTCTC 3635
E15mm ACTACACCCCTATTGAGGGGACGCCGTACTGCTTTTCCCGAAATGACTCCTTGAGTTCTC 3635
E15fr ACTACACGCCCATTGAAGGCACGCCGTGCTGTTTCTCCCGTAACGATTCACTGAGCTCAC 2636
```

```
      * ***** ** ***** ** ** ** * *** ** ** ** ** ** **  **** ** *

E15hs TAGATTTTGATGATGATGATGTTGACCTTTCCAGGGAAAAGGCTGAATTAAGAAAGGCAA 3695
E15mm TGGATTTTGATGATGACGATGTTGACCTTTCCAGGGAAAAGGCCGAGTTAAGAAAGGGCA 3695
E15fr TCGACTTCGATGAAGACGACGGTG---------GGGATAAAGACGAA--GAGCATAAGAA 2685
      * ** ** ***** ** ** * **         **** ** *  **    ** *     *

E15hs AAGAAAATAAGGAATCAGAGGCTAAAGTTACCAGCCACACAGAACTAACCTCCAACCAAC 3755
E15mm AAGAAAGCAAGGATTCCGAAGCCAAAGTTACCTGCCGCCCAGAACCAAACTCAAGCCAGC 3755
E15fr AACAAAAGAAGAAGATGGGAAGAAGAGGAAGCAGCAAACGGCGGCCGTCTTCCCTCGGAC 2745
      ** ***  *** *    *     * **  * * **         *     **   *   *

E15hs AATCAGCTAATAAGACACAAGCTATTGCAAAGCAGCCAATAA--ATCGAGGTCAGCCTAA 3813
E15mm AGGCAGCTAGTAAGTCACAAGCCAGTATAAAACATCCAGCAA--ACAGAGCACAGTCCAA 3813
E15fr AAA-GGCCAACGCGACGTCGTCGGATGAAAAGCAGAAGTTCGCCATCGAGGACACACCCG 2804
      *    ** *    * *     *   *  *** **          *  ***  **  *

E15hs ACCCATACTTCAGAAACAATCCACTTTTCCCCAGTCATCCAAAGACATACCAGACAGAGG 3873
E15mm ACCAGTGCTGCAGAAACAGCCCACTTTCCCCCAGTCCTCCAAAGACGGACCAGATAGAGG 3873
E15fr TCTGCTTTTCCAGAAAC------TCCTCCCTCAGCTCTCTGAGCG-ACATTGACCAGGAG 2857
       *   *  * *******         * ** ***   **  *     *      **  *

E15hs GGCAGCAACTGATGAAAAGTTACAGAATTTTGCTATTGAAAATACTCCGGTTTGCTTTTC 3933
E15mm GGCAGCAACTGACGAAAAACTGCAGAATTTTGCTATTGAAAATACTCCAGTTTGCTTTTC 3933
E15fr AATAACAACA-------------AGGAGTTTGC-------------------------C 2878
        * ****              ** * *****                          *

E15hs TCATAATTCCTCTCTGAGTTCTCTCAGTGACATTGACCAAGAAAACAACAATAA---AGA 3990
E15mm TCGAAATTCCTCTCTGAGTTCCCTTAGTGACATTGACCAGGAAAACAACAATAACAAAGA 3993
E15fr CCGCCGCCACCCGCTGAG------CAGCAAGATGGA--GGCAAAGCGGTG---------- 2920
       *       * * *****       **  * ** **     *** *

E15hs AAATGAACCTATCAAAGAGACTGAGCCCCCTGACTCACAGGGAGAACCAAGTAAACCTCA 4050
E15mm AAGTGAACCAATCAAAGAAGCTGAACCTGCCAACTCACAAGGAGAGCCCAGTAAGCCTCA 4053
E15fr AAATCGTCTCCTC-------CTCGACGGGCAGAATCAAAGCCACGGCCC------CCTGC 2967
      ** *   *   **       **   *   *  * *** *   *   **       ***
```

```
E15hs AGCATCAGGCTATGCTCCTAAATCATTTCATGTTGAAGATACCCCAGTTTGTTTCTCAAG 4110
E15mm GGCATCCGGGTATGCTCCCAAGTCCTTCCACGTCGAAGACACCCCTGTCTGTTTCTCAAG 4113
E15fr AGCTTCTGGCTATGCCCCAAAAGCCTTTCACGTGGAGGACACACCCGTCTGCTTCTCCAG 3027
       ** ** ** ***** ** **  * ** ** ** ** ** ** ** ** ** ***** **

E15hs AAACAGTTCTCTCAGTTCTCTTAGTATTGACTCTGAAGATGACCTGTTGCAGGAATGTAT 4170
E15mm AAACAGCTCTCTCAGTTCTCTTAGCATTGACTCTGAGGACGACCTGTTACAGGAGTGTAT 4173
E15fr GAACTCGTCGCTCAGCTCGTTAAGCATCGACTCGGAAGACGACCTGCTGCAGGAGTGCAT 3087
       ***   ** ***** **  * ** ** ***** ** ** ****** * ***** ** **

E15hs AAGCTCCGCAATGCCAAAAAAGAAAAAG-------CCTTCAAGAC-------TCAAGGGT 4216
E15mm AAGTTCTGCCATGCCAAAAAAGAAAAGG-------CCTTCAAGAC-------TCAAGAGT 4219
E15fr CAGTTCGGCCATGCCCAAGAAGAAGAAGAAAGCCGCTGCCAGCGCCACGCCATCCACTGT 3147
       ** ** ** ***** ** ***** * *       *   **   *       ** *  **

E15hs GATAATGAAAA-ACATAGTCCCAGAAATATGGGTGGCATATTAGGTGAAGATCTGACACT 4275
E15mm GAGAGCGAAAA-GCAGAGCCCTAGAAAAGTGGGTGGCATATTAGCTGAAGACCTGACGCT 4278
E15fr GGCAGCACCACCAGCTGCTCCCAAAGCTGAGAACAGCATTCTGGCTGAAGA---GGAGCC 3204
       *  *     *         ** * *     *    ****  * * ******   *   *

E15hs TGATTTGAAAGATATACAGAGACCAGATTCAGAACATGGTCTATCCCCTGATTCAGAAAA 4335
E15mm TGATTTGAAAGATCTACAGAGGCCAGATTCAGAACACGCTTTCTCCCCCGACTCAGAAAA 4338
E15fr C--CCGGAGATGCCTTCAGAGGTG----CCGAGAAGCCCCGCCTCTCCCGACTCGGAGTC 3258
         ** *    * *****        *   *         ** ** ** ** **

E15hs TTTTGATTGGAAAGCTATTCAGGAAGGTGCAAATTCCATAGTAAGTAGTTTACATCAAGC 4395
E15mm TTTTGACTGGAAAGCTATTCAGGAAGGCGCAAACTCCATAGTAAGTAGTTTGCACCAAGC 4398
E15fr CTTTGATTGGAAGGCCATTCAGGAAGGAGCCAACTCCATCGTCAGTAGCCTTAATGCCGC 3318
       ***** ***** ** ********** ** ** ***** ** *****  *  *    **

E15hs TGCTGC---TGCTGCATGTTTATCTAGACAAGCTTCGTCTGATTCAGATTCCATCCTTTC 4452
E15mm TGCTGCAGCCGCCGCGTGCTTATCTAGACAAGCGTCATCCGACTCAGATTCCATTCTGTC 4458
E15fr CGCTGCCGCCGCCACGTCCTTGTCCCGCCAGGCGTCATCAGACTCTGACTCTGTCCTGTC 3378
       *****    **  * *  ** **  * ** ** ** ** ** ** ** ** **  * ** **
```

```
E15hs CCTGAAATCAGG--------AATCTC-TCTGGGATCACCATTTCATCTTACACCTGAT-- 4501
E15mm ACTAAAGTCCGG--------CATTTC-TCTGGGATCGCCTTTTCATCTTACACCTGAT-- 4507
E15fr CCTCAAATCTGTGGGCTCACCATTTCATCTGCCATCAGCCAATAATAATGCAGAAGAGGA 3438
       ** ** ** *          ** ** ****  ***  *   * **  * **   **

E15hs CAAGAAGAAAAACCCTTTACAAGTAATAA-------AGGCCCACGAATTCTAAAACCAGG 4554
E15mm CAAGAGGAAAAGCCATTCACAAGCAATAA-------AGGCCCAAGAATTCTCAAACCTGG 4560
E15fr CAAGGTGGATGAGGCTGAGGAGGTGGCGGTAAAGCGAGGAGCACGAATCCTCAAGGCTGG 3498
      ****  * *      *    * *             ***  ** **** ** **  * **

E15hs GGAGAAAAGTACATTGGAAACTAAAAAGATAGAATCTGAAAG------TAAAGGAATCAA 4608
E15mm AGAGAAAAGCACATTAGAAGCAAAAAAAATAGAATCTGAAAA------CAAAGGAATCAA 4614
E15fr GGAGCGCACTACGCTTGATGCTAAAAAAGAGGAGGACGAGGAAGAAGCCAAGGGAGTGAG 3558
       ***   *  **  * **  * *****    **    **          ** *** * *

E15hs AGGAGGAAAAAAAGTTTATAAAAGTTTGATTACTGGAAAAGTTCGATCTAATTCAGAAAT 4668
E15mm AGGCGGGAAAAAGGTTTATAAAAGCTTGATTACGGGAAAGATTCGCTCCAATTCAGAAAT 4674
E15fr AGGTGGCAAAAAGGTGTACAGGAGTCTCATTACAGGCAAAGTCAGGGCGGAGCCAGCAGC 3618
      *** ** ***** ** ** *  **  * ***** ** **  *  *  *  *  *** *

E15hs TTCAGGCCAAATGAAACAGCCCCTTCAAGCAAACATGCCTTCAATCTCTCGAGGCAGGAC 4728
E15mm TTCCAGCCAAATGAAACAACCCCTCCCGACAAACATGCCTTCAATCTCAAGAGGCAGGAC 4734
E15fr GAGGGGGCGGAGCA---AGCCCAGGGCGGCAGCCGTGGCGAAAGCGCCAGGAGGCGG--C 3673
          * *  *  *   * ***       **  * ** *   *    *  ***** *  *

E15hs AATGATTCATATTCCAGGAGTTCGAAATAGCTCCTCAAGTACAAGTCCT---GTTTCTAA 4785
E15mm GATGATTCACATCCCAGGGCTTCGGAATAGCTCCTCTAGTACAAGCCCT---GTCTCTAA 4791
E15fr GACGCTGCTGACCACGGAGG---GGCATCCTCTCGAGATTCCACGCCGTCGCGTTCGTCC 3730
       * * * *  *   * *      *  **     *   * * ** * * *   **   *

E15hs AAAAGGCCCACCCCTTAAGACTCCAGCCTCCAAAAGC--CCTAGTGAAGGTCAAACAGCC 4843
E15mm GAAAGGCCCACCCCTCAAGACTCCAGCCTCTAAAAGC--CCCAGTGAAGGGCCGGGAGCT 4849
E15fr AATGTAAATATTCAGAAAGGAGGGAAACTGTCGCAGCTTCCACGTGCAGCATCTCCAGGA 3790
       *       *  *   ***     *  **     ***  **  *** **       **

E15hs ACCACTTCTCCTAGAGGAGCCAAGCCATCTGTGAAATCAGAATTAAGCCCTGTTGCCAGG 4903
```

```
E15mm ACCACTTCTCCTCGAGGAACTAAGCCAGCAGGAAAGTCAGAGCTTAGCCCTATCACCAGG 4909
E15fr AGCGCATCCTCGACCTCATCTTCGGCCTCCAGAGCAGCAAAACAGAGCGGGGTGACCAAA 3850
      * * * **  *      * *   * *  *        ** *    ***    *  ***

E15hs CAGACATCCCAAATAGGTGGG-TCAAGTAAAGCACCTTCTAGATCAGGATCTAGAGATTC 4962
E15mm CAAACTTCCCAAATCAGTGGG-TCAAATAAGGGGTCTTCTAGATCAGGATCTAGAGACTC 4968
E15fr GGCAGCACCGGCACCAACGGGCTCCCGCGGAGCG-------AATCGGCATCCAGGGTGGG 3903
         *   **   *     *** **       *          *** * *** ** *

E15hs GACCCCTTCAAGACCTGCCCAGCAACCATTAAGTAGACCTATACAGTCTCCTGGCCGAAA 5022
E15mm CACTCCCTCAAGACCTACACAGCAACCATTAAGTAGGCCAATGCAGTCTCCAGGGCGAAA 5028
E15fr CGGGTCCGCGGGGGCCAAGAAGCAGAAGACGGAGCCCGAGAAGCCGGCCCTCGTTCGCCA 3963
           *  *  *  *     ****                 *  * * * * *  *  **  *

E15hs CTCAATTTCCCCTGGTAGAAATGGAATAAGTCCTCCTAACAAATTATCTCAACTTCCAAG 5082
E15mm CTCAATTTCCCCTGGTAGAAATGGAATAAGCCCTCCTAACAAACTGTCTCAGCTGCCCAG 5088
E15fr GTCAACATTCATC---AAAGAAGCGCCGAGCCCGAC------GCTGAAGAGGAAGCTGGA 4014
       ****  * *      * * * *     ** **  *        *         *

E15hs GACATCATCCCCTAGTACTGCTTCAACTAAGTCCTCAGGTTCTGGAAAAATGTCATATAC 5142
E15mm AACATCATCTCCCAGTACTGCTTCAACTAAGTCCTCCGGTTCTGGGAAAATGTCATATAC 5148
E15fr GGAATCGGCCCCGGCGGCACCGTCGGAGTCGCCCACCAGCCCCGACATCTTCCTGCCCTC 4074
         ***  * **     *  * **      * ** *  *  * *  *   *        *

E15hs ATCTCCAGGTAGACAGATGAGCCAACAGAACCTTACCAAACAAACAGGTTTATCCA-AGA 5201
E15mm ATCCCCAGGTAGACAGCTGAGCCAACAAAATCTTACCAAACAAGCAAGTTTATCCA-AGA 5207
E15fr AGCGGGCAGGAGGCA-CGACGTCAACCGCTCCCACTCCGAGAGCCCGTCGCGACCACAGG 4133
      * *     * ** **     * ****     *    *  * *  *        *** **

E15hs ATGCCAGTAGTATTCCAAGAAGTGAGTCTGCCTCCAAAGGACTAAATCAGATGAATAATG 5261
E15mm ATGCCAGCAGTATCCCCAGAAGTGAGTCGGCATCTAAAGGACTGAATCAGATGAGTAACG 5267
E15fr AAGCCA----CGTCATCGCGCTTCAGCC-GCACCGGCACG-TGGAAGCGGGAGAACAGCA 4187
      * ****      *         * ** * **  *   * *    ** * *  **  *

E15hs GTAATGGAGCCAATAAAAAGGTAGAACTTTCTAGAATGTCTTCAACTAAATCAAGTGGAA 5321
E15mm GCAATGGGTCAAATAAAAAGGTAGAACTTTCTAGAATGTCTTCAACTAAATCAAGTGGAA 5327
```

```
E15fr GCACGGGAGCGGGCGG--AGGGAGCGCCGGCAAACAC-TCCACGTCTTTACCGCGGGTGG 4244
      * *  **  *        *** **  *   * *  *  **  *  **  * *  * *

E15hs GTGAATCTGATAGATCAGAAAGACCTGTATTAGTACGCCAGTCAACTTTCATCAAAGAAG 5381
E15mm GTGAATCAGACAGATCAGAAAGGCCTGCATTAGTACGCCAGTCTACTTTCATCAAAGAAG 5387
E15fr GAACGTGGAAGAGGACGGGAAGCTCCTCCTCCGTG---CTGTCCGCGTCC-TCGGAGTC- 4299
      *    *   * **  * * ***  *    *  **    * ***  * * * **  **

E15hs CTCCAAGCCCAACCTTAAGAAGAAAATTGGAGGAATCTGCTTCATTTGAATCTCTTTCTC 5441
E15mm CCCCAAGCCCAACCCTGAGGAGGAAACTGGAGGAATCTGCCTCATTTGAATCCCTTTCTC 5447
E15fr CAGTGAGAAGGGGCGCAGCGAGGAGGACGGCTCTCTCCGCTCCAAGGGGA---CGTGGAG 4356
      *    **      *      ** *    **     ** **  **   * *   * *

E15hs CATCATCTAGACCAGCTTCTCCCACTAGGTCCCAGGCACAAACTCCAGTTTTAAGTCCTT 5501
E15mm CATCTTCTAGACCAGATTCTCCCACCAGGTCGCAGGCACAGACCCCAGTTTTAAGCCCTT 5507
E15fr GAAGACCAAGAGCAGCGGCGGCGATTCGTCAGCGGG-ACGGGGC--------GGGTCCAA 4407
       *    * *** ***   *  * *   *    * ** **                 * **

E15hs CCCTTCCTGATATGTCTCTATCCACACATTCGTCTGTTCAGGCTGGTGGATGGCGAAAAC 5561
E15mm CCCTTCCTGATATGTCTCTGTCCACACATCCATCTGTTCAGGCAGGTGGGTGGCGAAAGC 5567
E15fr CAAAGCGGAGGACGTGTGGGTCCGCCTGGAGGACTGTCCCGTC-AACAACCCCCGCTCCT 4466
      *    *     * ** *   *** *        **** * * *          **

E15hs TCCCACCTAATCTCAGTCCCACTATAGAGTATAATGATGGAAGACCAGCAAAGCGCCATG 5621
E15mm TCCCGCCTAATCTCAGCCCCACTATCGAGTATAATGACGGAAGGCCCACAAAACGGCATG 5627
E15fr CCTCTTCCTGTTCTGCCCGCTCTCCCACCTCTGGCAACGCCCCTCCCATCAT------TG 4520
       * *  *   *      * * **      * *    * *     **    *       **

E15hs ATA-TTGCACGGTCTCATTCTGAAAGTCCTTCTAGACTTCCAATCAATAGGTCAGGAACC 5680
E15mm ATA-TTGCACGCTCCCATTCTGAAAGTCCTTCCAGACTACCAATCAACCGGGCGGGAACC 5686
E15fr ATAGCCCCACGCCCTCA-----AAAATTCTTTCTTCCTCCTCATCTTCCT--------CC 4567
      ***    ****  * **     *** * ***     ** *  ***             **

E15hs TGGAAACGTGAGC-ACAGCAAACATTCATCATCCCTTCCTCGAGTAAGCACTTGGAGAAG 5739
E15mm TGGAAGCGTGAAC-ACAGCAAACATTCCTCGTCCCTTCCTCGAGTGAGTACTTGGAGAAG 5745
E15fr TCTAATCTTAATCTACGGCGGAGCTACGAAAGC-----CTCGA-TGACAAGCCGCCGCCG 4621
```

```
      *  ** * * * * ** **  *  * *     *     ***** * *  *   * * *

E15hs AACTGGAAGTTCATCTTCAATTCTTTCTGCTTCATCAG-AATCCAGTGAAAAAGCAAAAA 5798
E15mm AACTGGAAGCTCATCTTCTATTCTTTCTGCTTCATCAG-AGTCCAGTGAAAAAGCAAAAA 5804
E15fr --CCGGAG---CGTCCGCAGCAGCAGCAGCGCAATCAGCAGCGCAGCGGCGCCGTAGCAG 4676
        * ***    * **  *        * **   ***** *   *** *     * *  *

E15hs GTGAGGATGAAAAACATGTGAACTCTATTTCAGGAACCAAACAAAGTAAAGAAAACCAAG 5858
E15mm GTGAGGATGAAAGGCATGTGAGCTCCATGCCAGCACCCAGACAGATGAAGGAAAACCAGG 5864
E15fr CCCGGGTC---AGCCCCTTTAACTACACCCCGAGCCCCAGAAAGAGCAACGCCGAC---G 4730
      **     *   *   * * **  *   *     *** * * *  ** *   **      *

E15hs TATCCGCAAAAGGAACATGGAGAAAAATAAAAGAAAATGAATTTTCTCCCACAAATAGTA 5918
E15mm TGCCCACCAAAGGAACATGGAGGAAAATCAAGGAAAGTGACATTTCTCCCACAGGCATGG 5924
E15fr CCTCCTCGACCAGCACCACGCCGACCACTACGCCTTCATCGTCATCGGCGACGCCCCCGC 4790
        ** * *   * **   *   *  *  *     **  * **

E15hs CTTCTCAGACCGTTTCCTCAGGTGCTACAAATGGTGCTGAATCAAAGACTCTAATTTATC 5978
E15mm CTTCTCAGAGCGCTTCCTCAGGTGCTGCCAGTGGTGCTGAATCCAAGCCTCTGATCTATC 5984
E15fr GGCCTTCGCTCATCCCCACCCCCGTCACCAAAAAGCGCGAGCCAAAGGG---GGGCGA-- 4845
        **  *  *    ** *    *   * *        **  * ***          *

E15hs AAATGGCACCTGCTGTTTCTAAAACAGAGGATGTTTGGGTGAGAATTGAGGACTGTCCCA 6038
E15mm AGATGGCACCTCCTGTCTCTAAAACAGAGGATGTTTGGGTGAGAATTGAGGACTGCCCCA 6044
E15fr AGGCGGCGCCGGAGGGGGCGGGGCCAGCGGA-----GAGCGCGGCTCGTACATTGTGACT 4900
      *   *** **    *   *     *** ***     * * * *  * *   * **   *

E15hs TTAACAATCCTAGATCTGGAAGATCTCCCACAGGTAATACTCCCCCGGTGATTGACAGTG 6098
E15mm TTAACAACCCTAGATCTGGACGGTCCCCCACAGGCAACACCCCCCCAGTGATTGACAGTG 6104
E15fr T---------------------------------------------------------- 4901
      *

E15hs TTTCAGAAAAGGCAAATCCAAACATTAAAGATTCAAAAGATAATCAGGCAAAACAAAATG 6158
E15mm TTTCAGAGAAGGGAAGTTCAAGCATTAAAGATTCAAAAGACACCCATGGGAAACAGAGTG 6164
E15fr ------------------------------------------------------------
```

```
E15hs TGGGTAATGGCAGTGTTCCCATGCGTACCGTGGGTTTGGAAAATCGCCTGAACTCCTTTA 6218
E15mm TGGGCAGTGGCAGT---CCTGTGCAAACCGTGGGTCTGGAAACCCGCCTCAACTCCTTTG 6221
E15fr ------------------------------------------------------------


E15hs TTCAGGTGGATGCCCCTGACCAAAAAGGAACTGAGATAAAACCAGGACAAAATAATCCTG 6278
E15mm TTCAGGTAGAGGCCCCAGAACAGAAAGGAACTGAGGCAAAACCAGGACAGAGTAACCCAG 6281
E15fr ------------------------------------------------------------


E15hs TCCCTGTATCAGAGACTAATGAAAGTTCTATAGTGGAACGTACCCCATTCAGTTCTAGCA 6338
E15mm TCTCTATAGCAGAGACTGCTGAGACGTGTATAGCAGAGCGTACCCCTTTCAGTTCCAGTA 6341
E15fr ------------------------------------------------------------


E15hs GCTCAAGCAAACACAGTTCACCTAGTGGGACTGTTGCTGCCAGAGTGACTCCTTTTAATT 6398
E15mm GCTCCAGCAAGCACAGCTCACCTAGCGGGACTGTTGCTGCCAGAGTGACACCTTTTAATT 6401
E15fr ------------------------------------------------------------


E15hs ACAACCCAAGCCCTAGGAAAAGCAGCGCAGATAGCACTTCAGCTCGGCCATCTCAGATCC 6458
E15mm ACAACCCTAGCCCTAGGAAGAGCAGCGCAGACAGCACTTCAGCCCGGCCGTCTCAGATCC 6461
E15fr ------------------------------------------------------------


E15hs CAACTCCAGTGAATAACAACACAAAGAAGCGAGATTCCAAAACTGACAGCACAGAATCCA 6518
E15mm CTACGCCAGTGAGCACCAACACGAAGAAGAGAGATTCGAAGACTGACAGCACAGAATCCA 6521
E15fr ------------------------------------------------------------


E15hs GTGGAACCCAAAGTCCTAAGCGCCATTCTGGGTCTTACCTTGTGACATCTGTTTAA 6574
E15mm GTGGAGCCCAAAGTCCTAAACGCCATTCCGGGTCTTACCTCGTGACGTCTGTTTAA 6577
E15fr ------------------------------------------------------
```

Appendix, Section 3B

Apc2 nucleotide alignment

```
E1hs ATGGCGAGCTCCGTGGCGCCCTACGAGCAGCTGGTGAGGCAGGTGGAGGCCTTGAAGGCT 60
E1mm ATGACCAGCTCCATGGCCTCATATGAGCAGCTGGTGCGCCAGGTGGAGGCCCTGAAGGCC 60
E1fr ----------------------TACGATCAGCTTGCCCACCAGGTGGCGGCTCTCCGCAAG 39
                          ** ** ***** *     ******* ***  *

E1hs GAGAACAGCCACCTGAGGCAGGAGCTAAGGGACAACTCCAGCCACCTGTCCAAGCTGGAG 120
E1mm GAGAACACTCACTTAAGGCAGGAGCTGAGGGATAACTCTAGCCATCTCTCCAAGCTTGAG 120
E1fr GAAAACTGTCACCTGCGGAGGGAGCTGGAGGACAACTCCAACCAACTGTCCAAACTGGAG 99
      ** ***   *** *  **  ******   *** ***** * *** ** ***** ** ***

E1hs ACAGAGACGTCGGGCATGAAG 141
E1mm ACAGAGACTTCTGGAATGAAG 141
E1fr ACTGAGACCTTCGGCATGAAG 120
      ** ***** *  ** ******

E2hs GAGGTCCTGAAGCACCTACAGGGAAAACTGGAGCAGGAGGCCCGAGTGCTGGTGTCC 57
E2mm GAGGTCCTGAAACACCTCCAGGGCAAGCTGGAGCAGGAGGCGAGAGTGCTCGTGTCT 57
E2fr GAAGTTCTGAAGCAACTGCAGAGTAAACTGGAGCAGGAGGCCGGAACTCTGGCCTCA 57
      ** ** ***** ** ** *** * ** **************  **   ** *  **

E2hs TCGGGGCAGACGGAGGTGCTGGAGCAGCTGAAGG 91
E2mm TCCGGGCAGACAGAAGTGTTAGAGCAACTGAAAG 91
E2fr TCCGGGAGGAGCGATGTCCTGCACCAGCTCAAA- 90
      ** ***  **  ** **  *  * ** ** **

E8hs GTGGAGGTGGTCTTCTGGCTGTTGTCCATGTTGGCGACGCGCGACCAGGAGGAT 54
E8mm GTGGAGGTGGTGTTCTGGCTTCTATCTATGTTGGCAACGCGCGACCAGGAAGAT 54
E8fr GTGGAGATGGTGTTCTGGCTGCTGTCCATTCTTACTAACAGAGATAAGGATGAG 54
     ****** **** ********  * ** **  *  * *   * **  **** **
```

```
E8hs ACAGCGCGCACGCTGCTGGCCATGTCCAGCTCGCCCGAGAGCTGCGTGGCCATGCGCCGC 114
E8mm ACTGCGCGCACACTGCTGGCCATGTCCAGCTCTCCAGAGAGCTGTGTAGCCATGCGCCGC 114
E8fr ATGTCTCGGACGCTGCTGGCGTTGTCCAGCTCCCCAGACAGTTGTATCGCCATGAGGAAG 114
     *   * ** ** ********  ********** ** ** ** **  * ****** *

E8hs TCGGGCTGTCTGCCTCTGCTGCTGCAAATCCTCCACGGCA-------CCGAGGCCGCGGC 167
E8mm TCGGGCTGTCTGCCACTGCTGCTCCAGATCCTTCATGGCA-------CTGAGGCTGGGTC 167
E8fr TCTGGCTGTGTCCCCCTGCTGGTCCAGATCCTGCATGACGGATCAGGCGGCACCAGCAGC 174
     ** ****** * ** ****** * ** ***** ** * *        * *   * *   *

E8hs CGGGGGT----CGCGCCGGG-GCCCCAGGGGCACCGGGCGCCAAGGACGCACGCATGCGC 222
E8mm TGTGGGG----CGCGCAGGG-ATCCCCGGAGCGCCTGGTGCCAAAGATGCACGCATGCGC 222
E8fr CTTGAGCAGTCCGCGTCGTGTGCCACGGTGACAGGCTGCAGCCGGGAGGCCAAGTCCAGG 234
        * *     ****  * *   * * *   *     *   *   ** **        *

E8hs GCCAACGCGGCGCTGCACAACATCGTCTTCTCGCAGCCGGACCAGGGCCTGGCGCGCAAG 282
E8mm GCCAATGCGGCCCTGCACAACATCGTTTTCTCCCAGCCGGATCAGGGCCTGGCACGCAAG 282
E8fr GCAAGCGCCACCCTCCACAACATCATCTTCTGTCAGCAGGATGAGGGACAGGCCCGCCGT 294
     ** *  **  * ** ********* * ****  **** ***  **** * *** ***

E8hs GAGATGCGCGTCCTGCACGTGCTGGAGCAGATCCGGGCCTACTGCGAGACCTGCTGGGAC 342
E8mm GAGATGCGTGTGCTGCATGTGCTGGAGCAGATCCGAGCCTACTGCGAGACCTGCTGGGAC 342
E8fr GAAAAAAGGGTTCTGCTCCTGCTGGAGCAGATCCGGACACACTGTGACACTGGCTGGGAT 354
     ** *   * ** ****   ***************  *  **** ** **  *******

E8hs TGGCTGCAGGCCCGAGACGGCGGGCCCGAG-GGAGGTGGCGCCGGCAGCG 391
E8mm TGGCTTCAGGCGCGGGACAGCGGGACAGAA-AGTG--------------- 376
E8fr TGGATTGAGAACC-ACACAGTGACTTCTACTGGAACTAGAACCAACG--- 400
     *** *  **   *   ** * *      *   *

E9hs ---CCCCGATCCCCATCGAGCCGCAGATCTGCCAGGCCACCTGTGCTGTTATGAAGCTGT 57
E9mm ---CTCCTGTCCCAATCGAGCCGCAGATCTGCCAGGCTACCTGTGCAGTGATGAAGCTGT 57
E9fr ATGTCCCAGAACCGATGGACCCGCAGATCTGTCAGGCTGTTTGTGCCATCATGAAACTCT 60
          **    ** ** ** ********** *****    *****  * ***** ** *
```

```
E9hs CCTTTGATGAGGAGTACCGCCGTGCCATGAACGAGCTAG 96
E9mm CATTTGACGAAGAATACCGTCGGGCTATGAATGAGCTAG 96
E9fr CCTTTGAGGAGGAGTACCGACACGCCCTGAAAGAGTTGG 99
       ***** ** ** ***** *  **  **** *** * *

E10hs GTGGGCTGCAGGCCGTGGCAGAGCTGCTGCAGGTTGACTATGAGATGCACAAGATGACCC 60
E10mm GGGGGCTGCAGGCTGTGGCAGAACTACTGCAGGTGGACTATGAAATGCACAAGATGACCC 60
E10fr GTGGTCTACAGGTCATAGCTGAACTGATCCATCTGGACCAGGAAACGCACGGCATGCAGA 60
      * ** ** ****   * ** ** **  * **  * *** * ** * ****   ***

E10hs GGGACCCGCTGAACCTGGCGCTGCGCCGCTACGCGGGCATGACCCTCACCAACCTCACCT 120
E10mm GGGACCCACTCAACCTTGCCCTGCGCCGCTACGCTGGCATGACCCTCACCAACCTCACCT 120
E10fr ATGACCCCATTAACATGGCTCTGCGACGCTATGCTGGGATGGCAATGACTAATCTGACCT 120
       *****  * *** * ** ***** ***** ** ** *** *  * ** ** ** ****

E10hs TTGGGGACGTTGCCAACAAG 140
E10mm TTGGAGACGTTGCCAACAAG 140
E10fr ACGGTGACGTGGTCAACAAG 140
       ** ***** * *******

E11hs GCCACCCTGTGTGCGCGCCGCGGCTGCATGGAGGCCATCGTGGCCCAGCTGGCCTCCGAC 60
E11mm GCCACACTGTGTGCCCGCCGAGGCTGCATGGAAGCCATTGTGGCCCAGCTTGGCTCTGAG 60
E11fr GCCACTTTGTGCTCAAAGAGAAGCTGCCTTCAGGCTCTGGTGGCTCAGCTGGCCTCTGAC 60
      *****  ****  *     *  ***** *  * **  * ***** ***** * *** **

E11hs AGTGAGGAGCTCCACCAG---------------------------------------------- 78
E11mm AGCGAGGAGCTGCATCAG---------------------------------------------- 78
E11fr AGCGAGGAGCTCCACCAAGTGGTCTCCAGTATCCTACGAAACCTGTCCTGGAGGGCCGAC 120
      ** ******** ** **

E11hs ------------------------------------------------------------
E11mm ------------------------------------------------------------
E11fr ATCAACAGCAAGAGAGTCCTGCGTGATATTGGTTGTGTGTCTGCTCTGATGACCTGTGCC 180
```

```
E11hs ---------------
E11mm ---------------
E11fr CTACAGGCCACCAAG 195

E12hs ------------------------------------------------------------
E12mm ------------------------------------------------------------
E11fr GCCACTTTGTGCTCAAAGAGAAGCTGCCTTCAGGCTCTGGTGGCTCAGCTGGCCTCTGAC 60


E12hs ------------------GTGGTGTCCAGCATCCTTCGGAACTTGTCCTGGAGGGCCGAC 42
E12mm ------------------GTTGTTTCCAGTATTCTGCGTAATCTGTCATGGAGGGCAGAC 42
E11fr AGCGAGGAGCTCCACCAAGTGGTCTCCAGTATCCTACGAAACCTGTCCTGGAGGGCCGAC 120
                       ** ** ***** ** ** ** **  **** ******** ***

E12hs ATCAACAGCAAGAAGGTGCTGAGGGAGGCGGGCAGCGTGACTGCCCTGGTGCAGTGTGTC 102
E12mm ATCAACAGCAAGAAGGTGCTGAGGGAGGTTGGCAGCATGACCGCCTTGATGGAATGTGTG 102
E11fr ATCAACAGCAAGAGAGTCCTGCGTGATATTGGTTGTGTGTCTGCTCTGATGACCTGTGCC 180
      *************  ** *** * **    **  *  ** * **  ** **   ****

E12hs CTGCGGGCCACCAAG 117
E12mm CTGCGGGCCTCCAAG 117
E11fr CTACAGGCCACCAAG 195
      ** * **** *****

E13hs GAGTCCACCCTGAAGAGCGTGCTGAGCGCCCTGTGGAATCTGTCTGCACACAGCACAGAG 60
E13mm GAGTCGACCCTAAAGAGTGTGCTCAGTGCTCTGTGGAACCTGTCAGCACACAGCACAGAG 60
E13fr GAGTCGACGCTGAAGAGCATCCTGAGTGCTCTGTGGAACCTGTCAGCTCACAGCATAGAC 60
      ***** ** ** *****  * ** ** ** ******** ***** ** ******* ***

E13hs AACAAGGCGGCCATCTGCCAGGTGGATGGCGCCCTGGGCTTCCTGGTGAGCACCCTGACC 120
E13mm AACAAGGCAGCCATCTGCCAGGTAGATGGTGCACTGGGTTTCCTGGTGAGCACCCTCACA 120
E13fr AACAAGGTGGCCATCTGTTCCGTAGACGGAGCACTGGGCTTCCTGGTGAGCACACTTACA 120
      *******  ********    ** ** ** ** ***** ************** ** **
```

```
E13hs TACAAGTGTCAGAGCAACTCGCTGGCCATCATCGAGAGCGGCGGCGGCATCCTCCGCAAT 180
E13mm TACCGTTGCCAAGGGAACTCCCTGGCAGTCATCGAGAGTGGCGGTGGGATCCTGCGCAAC 180
E13fr TACCGATGTCAGACCAACTCCCTGGCCATCATCGAGAGTGGCGGTGGGATCCTCCGAAAT 180
      ***   ** **    ***** *****  ********** ***** ** ***** ** **

E13hs GTGTCCAGCCTCGTCGCCACCCGTGAGGACTACAG 215
E13mm GTGTCAAGCCTCATTGCCACACGGGAGGACTACAG 215
E13fr GTGTCCAGCCTTGTTGCCACACGAGAAGACTAC-- 213
      ***** *****  * ***** ** ** ******

E14hs GCAGGTGCTCCGGGATCACAACTGTCTGCAGACGCTGCTGCAGCATCTGACTTCGCACAG 60
E14mm GCAGGTGCTCCGTGACCACAACTGCCTGCAGACACTGCTGCAGCACCTCACATCACACAG 60
E14fr ACAAATACTGCGGGATCACAACTGCCTCCAAACGCTGCTGCAGCACCTGCGCTCCCATAG 60
       **  * ** ** ** ******** ** ** ** ********** **    ** ** **

E14hs CCTGACCATCGTGAGCAACGCGTGCGGCACGCTCTGGAACCTGTCGGCCCGCAGCGCCCG 120
E14mm TTTGACCATCGTGAGCAATGCCTGTGGCACCCTCTGGAACCTGTCTGCCCGCAGCCCCCG 120
E14fr CCTGACTATAGTTAGTAATGCTTGTGGGACCCTCTGGAACCTGTCTGCAAGAAGTTCCAA 120
       **** ** ** ** ** ** ** ** ** ************** **  * **  **

E14hs TGACCAGGAGCTGCTGTGGGACCTGGGCGCCGTGGGCATGCTGCGTAATCTGGTGCACTC 180
E14mm CGATCAGGAACTGTTGTGGGACCTGGGGGCCGTGGGCATGCTACGCAACCTCGTCCACTC 180
E14fr AGACCAGGAACTGCTTTGGGAGCTGGGGGCTGTGAGCATGCTGCGCAACCTGATCCACTC 180
       ** ***** *** * ***** ***** ** *** ******* ** ** **  * *****

E14hs CAAGCACAAGATGATCGCCATGGGCAGCGCCGCCGCCCTGCGCAACCTGCTGGCCCATCG 240
E14mm CAAACACAAGATGATCGCCATGGGTAGTGCTGCTGCTCTGCGGAACCTGCTAGCCCACCG 240
E14fr CAAGCACAAGATGATAGCAATGGGCAGTGCTGCGGCCCTGAGAAACCTCCTGACCAATCG 240
      *** ********** ** ***** ** ** ** ** *** * ***** **  ** * **

E14hs GCCCGCCAAGCACCAGGCGGCCGCCACCGCCGTGTCCCCAGGCAGCTGCGTGCCCAGCCT 300
E14mm ACCCGCCAAGTATCAGGCCGCAGCCATGGCTGTCTCCCCAGGCACCTGCGTGCCCAGTCT 300
E14fr GCCCCTCAAGTACAAGGATACTGCC-----TGTGTCACCTGGTTCGTGCATGCCATCGCT 295
       ***  **** *  ***   * ***      ** ** ** **    *** ****    **
```

```
E14hs GTACGTGCGCAAGCAGCGGGCGCTGGAGGCCGAGCTGGACGCACGGCACCTCGCGCAGGC 360
E14mm TTACGTCCGCAAGCAGAGGGCTCTGGAAGCTGAGTTGGACACTCGGCACCTGGTGCATGC 360
E14fr CTACATGAGGAAACAAAAGGCCCTGGAAGCTGAGCTGGATGCCAAACACCTGGCAGAGAC 355
       *** *  * ** **   *** ***** ** *** ****  *    ***** *   *  *

E14hs GCTGGAGCACCTGGAGAAGCAGGGCCCGCCGGCAGCCGAGGCCGCCACTAAGAAGCCGCT 420
E14mm ACTCGGTCACTTAGAGAAGCAGAGTCTGCCTGAGGCAGAGACCACTTCAAAGAAGCCCCT 420
E14fr TTTTGACATAATCGAGAGACAGAATCC-CAGGCAACTGA-----CCCTCAACAGGCCACT 409
       * *      * ****  ***   *  *  *   * **     *    ** * *** **

E14hs GCCGCCCCTGCGACACCTGGACGGCCTGGCCCAAGACTATGCTTCCGATTCGGGCTGCTT 480
E14mm GCCACCCCTCCGCCACCTGGACGGGCTGGTGCAGGACTATGCCTCTGATTCTGGCTGCTT 480
E14fr AC---------GGCACATCGAGAGTCTAGCAAAGGATTACGCGTCTGATTCAGGCTGTTT 460
       *         * *** * **  * ** *   * ** ** ** ** ***** ***** **

E14hs TGACGACGACGATGCACCGTCATCCCTGGCTGCGGCCGCGGCCACCGGGGAGCCAGCCAG 540
E14mm TGACGATGATGATGCACC---ATCCCTGGCTGCTGCTGCCACCACAGCTGAGCCCGCCAG 537
E14fr TGATGACGATGAAGCTCC---TAGCGTGCCTAGTAATCTGGACACAGGG---------AG 508
      *** ** ** ** ** **      * ** ** *** * **

E14hs CCCTGCCGCGCTGTCCCTCTTCCTGGGCAGCCC---CTTCCTGCAGGGGCAG--GCGCTG 595
E14mm CCCAGCAGTGATGTCTATGTTCCTTGGCGGTCC---CTTCCTTCAGGGCCAG--GCACTG 592
E14fr CTTCTCTATGCTGTCTATGTTTCTTACCAATTCTAACTTCTCACAGAACCAACAACGCAA 568
      *    *   * ****  * ** **   *    *   ****   ***   **    * *

E14hs GCTCG-CACCCCGCCCACCCGCCGAGGCGGCAAGGAGGCAGAGAAGGACACCAGTGGGGA 654
E14mm GCCCG-CACCCCACCTGCCCGCCAGGGTGGCCTAGAAGCCGAGAAGGAGGCTGGTGGGGA 651
E14fr AAGGGACAATGAACCCGAGAGAGACGGAGATTCACCAGCAAGTGGCGAGAACAAATATCC 628
         * **     **     *    ** *        **       **

E14hs GGCAGCCGTGGCGGCCAAGGCCAAGGCCAAGCTGGCGCTTG------------CAGTGGC 702
E14mm GGCAGCTGTGGCTGCCAAGGCCAAGGCCAAGCTGGCGTTGG------------CTGTGGC 699
E14fr AGCCGATGCCGTATCTGCTGCTGCAGATAAACTGGCGCAGAAAATCACCAACACTGTGGC 688
       ** *  *  *   *    **    *  ** ******               * *****
```

```
E14hs GCGCATCGACCAGCTGGTGGAGGACATCTCCGCCCTGCACACCTCGTCCGACGATAGCTT 762
E14mm TCGGATCGACAGATTGGTGGAGGACATCTCTGCCCTGCACACCTCATCAGACGACAGCTT 759
E14fr CAAGATTGACAGATTGGTGGATGACATTACTA---TGCACACGTCTTCTGAGGACAGTTT 745
         ** ***    ******* *****  *      ******* ** ** ** ** ** **

E14hs CAGCCTCAGCTCTGGAGACC-CGGGACAGG--------AGGCGCCACGGGAGGGCCGCGC 813
E14mm CAGTCTCAGCTCGGGGGACC-CTGGGCAGG--------AGGCGCCAAGGGAGGGCCGTGC 810
E14fr CAGCCTGTGCTCTGAGGACCACCTGGCAGATTGGCCTTACGGACCACATGAGCTCAACGA 805
      *** **  **** *  **** *  * ***         * *  ***   ***  *   *

E14hs CCAG--------TCCTGCTCGCCATGCC-------GCGGCCCGGAGGGCGGGCGGCG-AG 857
E14mm TCAG--------TCCTGCTCTCCATGCC-------GGGGCACTGAGGGTGGGAGGCG-TG 854
E14fr CCAGACCAAGAGCCCTTCCCTCCTGTCCCATCTATGTGACACCAGCAGTGTGGTCCACAG 865
      ***         *** * * **    **       * * * *     * * *   *   *

E14hs AGGCAGG--AAGCCGGGCGCACCCGCTGCTGCGGCTCAAGGCGGCCCACGCCAGCCTCTC 915
E14mm AGGCTGG--CAGCCGGGCGCACCCTCTTTTGAGGCTCAAGGCAGCCCACACTAGCCTCTC 912
E14fr AGACCGCTTCAGCAGAGCTCGTGCACTCTTGCGTCTGAAAACGGCACAGTCGAGTTTATC 925
      ** * *    *** * ** *   * **  ** * ** **  * ** **  * **  * **

E14hs CAACGACAGCCTCAACAGCGGCAGTGCCAGCGACGGGTACTGCCCACGCGAACATATGCT 975
E14mm TAATGATAGCTTGAACAGCGGTAGCACCAGCGATGGCTACTGTACCCGGGAACACATGAC 972
E14fr GACAGACAGTCTGAACAGTGGTAGCACAAGTGATGGCTACTGTGGCAGCAAAGACCAGCT 985
       *  ** **  * ***** ** **  * ** ** ** *****     *  ** *   *

E14hs GC--CCTG----------CCCGCTGG-----CCGC----------ACTGG---------- 998
E14mm GC--CTTG----------CCCGCTGG-----CTGC----------ACTGG---------- 995
E14fr GCAACCTGTTACAAGAGCCCTGATGATGCAACAGCGACCCAAGCAACTGGATCTCAAATT 1045
      **  * **          ** * **      * **          *****

E14hs ----------------------------CTTCGCGCCGCGAGGACCCCAGGTGTGGGCA 1029
E14mm ----------------------------CCGAGCACCGTGATGACCCTGTGCGCGGACA 1026
E14fr AGCCCACAAAGATTACCAAGAAGCTGACTCTTCTCTTTGTAGTAACACTAATCGCCAGGA 1105
                                  *    *   *     ** *     *     *
```

```
E14hs GCCTC--------GGCCCAGCCGGC--TTGACCTTGACCTGCC-CGGCTGC--------- 1069
E14mm GACTC--------GGCCCAGCCGAC--TGGACCTGGACCTTCC-CAGCCG---------- 1065
E14fr GATTCCTGAGAGAGATTCTGTGGACAACAAACATTCACCAGTTGCAGATGGAGGCAAAAA 1165
      *  **        *   * *  * *     ** *  ***     * *  *

E14hs -CAGGCCGAGCCC---------CCGGCCCGCGAGGCCACC------TCCGCCGACGCCCG 1113
E14mm ---GGCTGAGCTT---------CCTGCCCGGGACACAGCA------GCCACCGATGCCCG 1107
E14fr GCAGGTTCAGCCTGTCCAGACACCTGTCAGTGCGTCCACCAAAGTGTCCTCGGATGTCAG 1225
        **   ***           ** * * * *   *  *        ** * ** * * *

E14hs CGTGCGCACCATCAAGCTGTCGCCTACCTATCAGCACGTGCCACTGCTTGAGGGTGCCTC 1173
E14mm AGTGCGCACAATCAAGTTATCCCCAACCTATCAGCATGTTCCACTGCTGGATGGGGCCGC 1167
E14fr CATGGCATCAATAAAACTGTCTCCTTCTTACCAGCAAGTCCCCAGTATTCAGAGTGTGGC 1285
        **    * ** **  * ** **  * ** ***** ** **     *  *  * *   *

E14hs AAGGGCGGGTGCAGAG----CCCCTCGCGGGGCCTGGAATCTCTCCAGGGGCCCGGAAGC 1229
E14mm TGGGGCAGGTGTTCGA----CCCCTGGTTGGGCCGGGAACCTCCCCGGGGGCTCGGAAAC 1223
E14fr CAAATTTGGTGTGGAAAAGGCCAGCGGTGATGTCCAGA-CAGCCCAGGCGATGAGGAAAC 1344
             ****         **    *    * *  **    * *  * *    **** *

E14hs AGGCCTGGCTGCCGGCAG--ACCACCTGAGCAAGGTTCCCGAGAA-------GCTGGCGG 1280
E14mm AGGCATGGATACCTGCAG--ACAGCCTGAGCAAAGTCCCTGAGAA-------ACTGGTGG 1274
E14fr AGCCCTCGGTCCCCACTGCCATGACAGCAGCGAGCATTACAAAAGCTCCTACATTAGCCT 1404
      ** * * * * **  * *  *   *   *** *        * *          * *

E14hs CTGCC------CCGCTGTCTGTGGCCAGCAAGGCACTGCAGAAACTGGCGGCGCAAGAGG 1334
E14mm CCTCT------CCACTGCCCATAGCTAGCAAGGTGCTGCAGAAGCTGGTCGCACAGGATG 1328
E14fr CCACCAAAAGCCCGAACATTGGGACGATGGAGACGGTCCAGAAGTACTCAGTGGAGAACA 1464
      *  *       **          * *   **    * *****       *   *  *

E14hs GGCCACTCTCGCTGTCCCGATGCAGCTCCCTTTCCTCGCTGTCCTCGGCCGGCCGCCCAG 1394
E14mm GGCCGATGTCCCTCTCCAGGTGCAGCTCTCTGTCCTCTCTATCTTCCACGGGCCATGCTG 1388
E14fr CCCCCATCTGCTTCTCCCGCTGTAGTTCTCTTTCCTCTCTTTCTTCCGGTGATGGAGCGC 1524
        **  * *   * *** * ** ** ** ** ***** ** ** **    *     *
```

```
E14hs GCCCCAGCGAGGGTGGTGACCTGGATGACAGTGACTCCTCCCTGGAGGGGCTGGA---GG 1451
E14mm TCCCCAGCCAGGCAGAGAACCT---TGACAGCGATTCATCCCTGGAGGGGCTTGA---GG 1442
E14fr TGGATGCACAAAGTGATAATGAGATGGAGAGCGACTCCTCGTTAGAAATAATTGATGTGG 1584
                *    *   *       ** ** ** ** **  * **     * **   **

E14hs AGGCCG--GCCCCAGCGA----GGCTGAG----------------------CTG---GACA 1481
E14mm AGGCTG--GTCCTGGTGA----GGCCGAG----------------------CTG---GGCA 1472
E14fr AGGATGAAGACTTGGTGAAAAAGGCTGAGGAGGATGAAACCTTGGAAGACCTGACCGATA 1644
      ***  *  * *   * **    *** ***                       ***   *  *

E14hs GCACGTGGCGGGCGCCCGGGGCCACCTCGCTGCCC-GTAGCCATTCCGGCTCCC-CGGCG 1539
E14mm GGGCGTGGCGAGCATCCGGGTCCACCTCTCTTCCG-GTGTCCATCCCAGCCCCG-CAGCG 1530
E14fr GCCTGCTGCTGATGACCGACTCCAAATCCTTTCCCAGTAAAGACACTGGTCCCGTCAGTA 1704
      *   *  **      ***   ***  **  * **  **    *  *  *  **  * *

E14hs TAACC------------GAGGCCGGGGC----CTGGGGG------------------TGG 1565
E14mm GGGGC------------GCAGTAGAGGT----CTTGGGG------------------TGG 1556
E14fr AGACCTGCTCTTCTAAAAAGGAAAAGGTGTTCCTTAGAGCAGTTTCCCCAGCCATAGTGG 1764
         *               *    **     **  * *                  ***

E14hs AAGACGCCACGCCGTCCAGCTCGTCGGAGAACTACGTGCAGGAGACACCGCTTGTGCTGA 1625
E14mm AGGATGCAACACCATCCAGCTCATCTGAGAACTGCGTCCAGGAGACGCCTTTGGTCCTGA 1616
E14fr AAGACCAATCTCCTTCAAGTTCATCTGAACACTACATACATGAGACCCCCCTGGTCATGA 1824
      * **     * ** ** ** ** ** ** **  *** * * ** ***** **  * **  ***

E14hs GCCGCTGCAGCTCTGTGAGCTCGCTGGGCAGCTTCGAGAGCCCGTCCATCGCCAGCTCCA 1685
E14mm GCCGTTGTAGTTCCGTGAGCTCCCTGGGCAGCTTCGAGAGCCGCTCCATTGCCAGCTCCA 1676
E14fr GCCGCTGCAGCTCGGTTAGCTCGTTAGGCAGCTTTGAGTCTCCCTCTATTGTCAGCTCAA 1884
      **** ** ** ** ** *****  * ******** ***   *  ** ** * ****** *

E14hs TCCCCAGTGAACCTTGCAGCGGGCAGGGCAGCGGCACCATCAGCCCTAGCGAGCTGCCCG 1745
E14mm TCCCCAGTGACCCGTGCAGCGGGCTGGGCAGTGGCACAGTGAGTCCCAGCGAGCTGCCAG 1736
E14fr TCCAGAGCGATCCCTGCAGCGAGATGATCGATGGAACAATAAGCCCGAGTGATCTTCCCG 1944
      ***  ** ** ** ******* *  *  *   ** **  * ** ** ** ** ** ** * *
```

```
E14hs ACAGCCCCGGACAGACCATGCCTCCCAGCCGGAGCAAGACGCCACCGCTGGCGCCCGCGC 1805
E14mm ACAGCCCCGGGCAGACGATGCCACCGAGCCGCAGCAAGACGCCACCG---------GCAC 1787
E14fr ACAGCCCTGGACAGACCATGCCTCCCAGTCGAAGTAAGACGCCATGTTGC---CTTGAGT 2001
      ******* ** ***** ***** ** ** ** ** *********           *

E14hs CACAGGGTCCCCC--CGAGGCCACCCAGTTCAGCCTGC-AGTGGGAGAGCTACGTGAAGC 1862
E14mm CTCCTGGGCAGCC--TGAGACCAGCCAGTTCAGCCTGC-AGTGGGAAAGCTATGTGAAGC 1844
E14fr CACATGTTCAAGAAACACAGACGGCAGGAATAGTCAGCCAATGGGAAGGCAGCCTGCGGA 2061
      * *  *  *           *  *  *   ** * ** * *****  **    **  *

E14hs GCTTCCTGGACATCGCCGACTGCCGGGAGCGCTGCCGGCTGCCATCTGAGCTGGACGCAG 1922
E14mm GCTTCCTAGACATCGCAGACTGTCGAGAAAGATGCCAGCCGCCCTCGGAGCTGGACGCGG 1904
E14fr CGTTCATGGAAATTGCAGACTCAAAGGACCGACTGAGCCTCCCCCCTGACCTGGACACA- 2120
       *** * ** ** ** ****     **  *       *  **  * ** ****** *

E14hs GCAGCGTGCGCTTTACCGTGGAGAAGCCAGACGAGAACTTCTCGTGCGCCTCCAGCCTCA 1982
E14mm GCAGCGTTCGCTTCACAGTGGAGAAGCCAGATGAGAATTTCTCCTGCGCCTCCAGCCTCA 1964
E14fr --ATGATTTACTTCACGGTGGAAAAGCCTACTGAAAACTTCTCGTGCGCTTCAAGTTTGA 2178
       *   *   *** ** ***** *****    ** ** ***** ***** ** **  * *

E14hs GCGCGCTGGCCTTGCACGAGCACTACGTGCAGCAGGACGTGGAGCTGCGGCTGCTGCCCT 2042
E14mm GTGCACTGGCCCTTCATGAGCTGTATGTCCAGCAGGATGTGGAGCTGCGTCTGAGGCCAC 2024
E14fr GTGCCTTGCCTCTTCATGAACACTACATACAGAAGGACGTCGAGCTAAAGTTAACACCAC 2238
      * **  ** *  * ** ** *  **  * *** **** ** *****     *    **

E14hs CGGCC--------TGCCCCGAGCGCGGCGGG----GGCGCCGGGGGCGCCGGCCTCCACT 2090
E14mm CTGCC--------TGCCCAGAGCGTGCGGTG----GGTGGTGGGGGCCACCGTCGGAGGG 2072
E14fr TGCTCCAGCAAGGTGACAAAAACCTTTTGTGTCATGATGAAGAGGGCCAGGAATTTGACA 2298
          *        ** *   * *     * *    *  *  * ****

E14hs TTGCAGGGCACCGGCGGCGGGAGGAGGGGCCGGCGCCCACG----GGTTCTC-------G 2139
E14mm ACG-AGGCTGCCAGCCGCCTG--GATGGTCCAGCACCAGCT----GGTTCTA-------G 2118
E14fr ATGGGGAGCGATA-CAGTGAG-GGAAACTCCGATGATGATATTGAAATCCTAAAAGAATG 2356
       *  *        * *   *  **    **               * **        *
```

```
E14hs CCCTCGCGGCGCCG-------CGGACCAGGAGCTGG-AACTGCTGCGGGAGTGCCTGGGA 2191
E14mm GGCTCGGTCAGCAA-------CTGATAAAGAACTGG-AGGCTTTGCGTGAATGCCTGGGG 2170
E14fr TATTAACTCAGCAATGCCCTCCAAATTTAGAAAAGTTAGACCTTCCCTGATGACCCAAAT 2416
         *      **         *  *    **   *  *     * *  **   **

E14hs GCCGCC-------GTGCCTGCCCGGCTGCGCAAGGTGGCC----TCCGCGCTGGTGCCAG 2240
E14mm GCAGCC-------ATGCCCGCCCGGCTCCGCAAGGTGGCC----TCGGCCTTGGTGCCTG 2219
E14fr CCCTCCTCATCTTGTGAGCTCTCAGACCCATAAACCAATCCATCTCCCAGTTTACATGAT 2476
       *  **        **    * * *   *  **      *    **     *

E14hs GTCGCCGC---GCACTCCCCGTGCCCGTCTA--CATGTTGGTGCCCGCCCC------GGC 2289
E14mm GCCGCCGC---TCATTGCCAGTCCCTGTGTA--CATGTTAGTGCCCGCCCC------GGC 2268
E14fr GTTCCCGAATGGCAAAACCCAAATCTGTCCGGGCAGGAAAGTCATCGTTTCACACAAGGA 2536
       *   ***     **   **     * **     ** *   **   **   *      **

E14hs CCCGGCCCAGGAGGACGACTCCTGCACTGACTCCGCGGAGGGCACGCCGGTCAACTTCTC 2349
E14mm TC------GGGGTGATGACTCGGGCACGGACTCCGCAGAGGGCACACCCGTTAACTTTTC 2322
E14fr TCTGAAATTAGACGATTCGTCCCTCACAGATTCCGCTGAAGGAACACCTGGTAACTTCTC 2596
       *        *  **    **   *** ** ***** ** ** ** ** *  ***** **

E14hs TAGCGCCGCCTCGCTCAGCGACGAGACGCTGCAGGGACCCCCCAGGGACCAGCC-CGGGG 2408
E14mm CAGTGCAGCCTCGCTCAGTGACGAGACCTTACAGGGACCCTCCAGGGACAAGCC-AGCCG 2381
E14fr CACAACAACATCGCTAAGTGACGACACGCTTCAGTATCCAGTGAAGCACAGGGGGAGCAA 2656
       *   *  * ***** ** ***** **  * ***   **    * * **  *    *

E14hs GACCAGCGGGCAGGCAAAGACCCACCGGCCGCCCCACCTCTGCCAGACAGGCCATGGGGC 2468
E14mm GGCCTGGGGACAGGCAGAAACCTACAGGCCGAGCTGCCCCTGCCAGGCAGACCCGATCTC 2441
E14fr AGACTGCTTATCTACCAGTATCATCAAG--GAGCAAGAGCTTGATGACGAAAAGAGGATC 2714
       * *        *    * *  *  *  *  *     **    * *           *

E14hs ACCGGCACAAGGCGGGAGGCGCCGGCCGCAGCGCGGAGCAGTCTCGGGGCGCGGGCAAGA 2528
E14mm ACCGGCCCAAAGCAGCAGGCGCTGGTAAGAGCACAGAACACACCCGGGGGCCCTGTAGGA 2501
E14fr GAAGATTTGAGGATATTCTCACACTTCCACAAACTGAACAGGACAAACAACCCTG---GA 2771
       *     * *       * *           * ** **           *  *   **
```

```
E14hs ACAGAGCAGGGCTGGAGCTGCCCCTGGGCCGGCCCCCGAGCGCCCCCGCAGAC---AAGG 2585
E14mm ACCGGGCAGGATTGGAGCTACCACTCAGCAGGCCCCAGAGTGCTCGGTCCAAC---AGGG 2558
E14fr ACACAGAAAAACCGACAC-ATCACTC--CGACTCAGAGGGTGCTGATGCAAAGCAAAGAG 2828
      **   * *     *   *   * **   *    *   * * **     * *    *  *

E14hs ACGGCTCAAAGCCCGGCCGGACCCGCGGGGACGGGGCGCTCCAGTCGCTGTGCCTCACGA 2645
E14mm ATAGCTCATGCCAGACCCGGACCCGCGGAGATGGTGCCCTGCAGTCGCTATGCCTCACAA 2618
E14fr GTAGCTGACCGAGTGGCTAG-CCAAAGGAATCGCGATCGGTCTCCCAACCAACAAAAGAA 2887
         *** *        *  * **   **    *        *   *      *   *  *

E14hs CGCCCACTGAGGAGGCCGTGTACTGCTTCTA-CGGCAACGACTCGGACGAGGAGCCCCCG 2704
E14mm CACCCACAGAGGAAGCTGTGTACTGCTTCTA-TG------ACTCTGATGAGGAACCACCA 2671
E14fr CAGACCGTGTCAAGATCTTGTTCGGAATCTGGCGCTTCCTGTCATAATGAAAAGTGACCA 2947
      *   *   *   *     *** * *  ***   *            * **  *    **

E14hs GCGGCCGCGCCCACGCCAACCCACCGGCGCACATCGGCCATCCCTCGCGCTTTTACGCGG 2764
E14mm GCCACTGCACCACCACC------TCGGCGGGCATCCGCCATCCCACGGGCTCTAAAGCGC 2725
E14fr AGATGCAGGATTCCAGAAAAGAAACATAATATGTAAAAAGATTGCACATTTTCTACATGA 3007
                  *          *        *                 *   *   *

E14hs GAGCGTCCGCAGGGCCGGAAGGAGGCCCCTGCCCCGTCCAAGGCTGCACCAGCTGCCCCG 2824
E14mm GAGAAACCGGCAGGCAGAAAGGAGA------CTCCATCCAGGGCAGCCCAGCCTGCCACA 2779
E14fr CAATAACTATGTGTGTGAAGACAG-------CAATGCTAACAGCGACGCTTTTGGACAGA 3060
       *    *     *   * *   **       *       *  **  * *     * *

E14hs CCGCCCGCCCGGACCCAGCCCAGCCTCATTGCTGACGAGACCCCGCCCTGCTACTCCCTG 2884
E14mm CTCCCTGTGAGAGCCCAACCCAGACTGATCGTGGATGAGACCCCGCCCTGCTATTCCCTG 2839
E14fr CGACCCG----AAAACAGACCAG---GGAAGCAAATGCAGCCAATACTGGCAAGAAGGAG 3113
      *  ** *        **  ****       *   * *   **    *  ** *      *

E14hs AGCTCCTCCGCCAGCTCCCTCAGCGAGCCCGAGCCCTCGGAGCCGCCGGCCGTCCATCCA 2944
E14mm ACTTCCTCAGCTAGTTCCCTCAGTGAGCCTGAGGCCCCTGAACAGCCGGCCAACCATGCT 2899
E14fr AATTCCCAAGGAAAA---CACAGAGGCTTTCAGAGAATAAAACAAA-GTCTTCTCAAGGA 3169
      *  ***   *  *     * *** *      **       * *    * *    **
```

```
E14hs CGAGGCCGGGAGCCCGCGGTCACCAAGGACCCGGGCCCAGGAGGCGGACGCGACAGCTCG 3004
E14mm CGAGGCCCGGAGCA----------------------GGGTAGTAAACAGGACAGCTCT 2935
E14fr TGAATCTATGGAGGGTTATTCGCTCAGCTCCTCCCTCAGCTCACTGAGTGATGCTGAGTT 3229
       **  *   *                                          * *

E14hs CCCAGCCCGCGGGCCGCGGAGGAGCTTCTGCAGCGGTGCATCAGCTCGGCCCTGCCCAGG 3064
E14mm CCTAGCCCAAGGGCAGAAGAGGAACTTCTGCAGAGATGTATCAGCTTGGCCATGCCCAGG 2995
E14fr TGAGGC----GGGCAAATCAAAAGCCCAGCAAACGTGGTACAAAAACAGACAGAACAAAA 3285
         **    ****     *  * *      *  *  * *  *     * *    * *

E14hs CGCCGGCCCCCCGTGTCTGGCCTGCGGCGCC-GCAAGCCCCGAGCCACCCGGCTGGATGA 3123
E14mm CGCCGGACCCAGGTGCCTGGCTCACGGCGTC-GCAAGCCCAGAGCCTTGAGGTCAGACAT 3054
E14fr CGCTGAATGCG---GTCCAACAAACGAAGCCCGTGAGCATTCACAGTCAATATGAGGAGC 3342
      *** *    *    * *   *   **  * * *  ***    *             *

E14hs GCGGCCCGCAGAGGGGTCCCGGGAACGCGGCGAGGAGGCAGCGGGCTCGGACCGGGCCTC 3183
E14mm AAGGCCCACTGAGATAACCCAGAAATGCCAGGAGGAGGTGGCTGGCTCCGATCCAGCCTC 3114
E14fr CCAGCTCCACGAGCTCTGTCAGTA-TGGATTCAGAGGATGATCTCCTCCAGAAATGCATA 3401
        ** *   ***      * * *  *     **  *        ***       ** *

E14hs CGACCTG----GATAGCGTGGAGTGGCGCGCCATCCAGGAGGGCGCCAATTCAATTGTCA 3239
E14mm TGACCTA----GACAGTGTGGAGTGGCAGGCTATCCAAGAGGGCGCAAACTCCATTGTCA 3170
E14fr ACATCTGCGATGCCAAAGCAGAGAAGGAAGCACGCAGCAAGGAAGAAGAAAGCAATGAAT 3461
       * **     *  *  *  ***  *   **   *    ***  *   *    * **

E14hs CGTGGCTGCACCAGGCAGCAGCT---GCCACGCGGGAGGCCTCGTCCGAGTCCGACTCCA 3296
E14mm CATGGCTGCATCAGGCAGCGGCCAAAGCCAGCCTGGAAGCATCTTCTGAGTCTGACTCCC 3230
E14fr AGTGACAAAGGTAAGAAGGCCCT---GGATGCCTGGAAGC----TGGAGGAGGAATTAGA 3514
       ** *      * * **   *    *     * *** **    *    *    * *

E14hs TCCTGTCCTTCGTATCCGGGCTGTCAGTGGGATCCACCCTACAGCCCCCCAAGCACAGG- 3355
E14mm TCTTGTCTTTGGTGTCCGGGGTGTCAGCAGGCTCCACCCTCCAGCCCTCCAAGCTCAGG- 3289
E14fr TAGTGATGATGCAGATTCAGATCTCAACAGTGTTGAGTGGAGAGCTATCCAAGAAGGTGC 3574
      *  **    *         * * ***   *  *  *      ***   *****     *
```

```
E14hs --AAGGGACGACAGGCGGAGGGAGAAATGGGCAGTGCCCGGCGGCCAGAGAAAAGGGGCG 3413
E14mm --AAAGGGCGAAAGCCTGCAGCAGAGGCTGGAGGTGCCTGGCGTCCTGAGAAACGGGGCA 3347
E14fr CAACTGGTTGTCACTGGGCTGCAGGCCTCAAAATCTCAAGAGCCTTCCTCTGAAGAGACT 3634
        *  **  *  *    *  * **          *  *          * * * *

E14hs CAGCCTCAGTCAAGACCAGCGGGAGCCCCCGTTCCCCTGCAGGCCCCGAGAAGCCACGTG 3473
E14mm CAACTTCCACTAAGATCAATGGGAGTCCCCGGCTACCTAACGGTCCTGAGAAGGCAAAGG 3407
E14fr GAATCGGTCATTTCATTCATGTCAACCTCTAGCTTCACGCCTAAAGAAAGGAAGTTTTGC 3694
       *           *     *  *  * *     *           ** *

E14hs GCACACAGAAGACCACGCCCGGGGTGCCAGCTGTGCTC--CGGGGACGAACAGTGATCTA 3531
E14mm GTACCCAGAAAATGATGGCAGGGGAGTCAACCATGCTC--CGGGGACGGACAGTGATCTA 3465
E14fr AAAGACAAAAAGTCA-AACAAACCTCTAGACTTTGCTCAACGCAAGCCCGTCCCAAACTT 3753
        *  ** **    *   *          *  *****  **    *        * **

E14hs CGTCCCCA-GCCCGGCACCCCGTGCCCAGCCCAAAGGGACCCCCGGCCCCCGCGCCACAC 3590
E14mm CTCAGCCG-GCCCAGCCTCCCGCACTCAGTCCAAAGGTATTTCTGGACCTTGTACCACAC 3524
E14fr TCCAGTTGTGTTCAGAGGCAGGACTGTAATTTACACACCGAGAAAGGAGACA-GCCCCAT 3812
               *  * *   *  *     *    * *          *       ** **

E14hs CGCGGAAGGTGGCGCCCCCTTGCCTGGCACAGCCCGCGGCTCCAGCCAAAGTCCCGAGCC 3650
E14mm CTAAGAAGACAGGGACATCTGGTACCACTCAGCCAGAAACTGTCACCAAAGCCCCCAGCC 3584
E14fr CGCAAAGGCCTCCTCCCACCAAACTGACGACCAGCTCAGATCCTCCAAAAAACCCAAACC 3872
       *    * *       *  *        *            *    * ***  *** * **

E14hs CCGGGCAGCAGCGGTCGCGGAGCCTACACCGGCCTGCCAAGACCTCGGAGCTGGCGACGC 3710
E14mm CTGAGCAACAACGTTCACGGAGCCTCCACCGACCGGGCAAGATCTCTGAGCTGGCAGCTT 3644
E14fr TTGCTCAACACAGATCAAAGAGCCTTCATCGATTGGG-ACACTCCCAGGAC--ATGGACC 3929
        *  ** **  * **   ****** ** **    *  *    * * *  *

E14hs TGAGCCAGCCCCCCAGAAGCGCCACACCGCCCGCCCGCCTCGCCAAGACCCCCTCCTCCA 3770
E14mm TGCGCCACCCGCCCAGGAGTGCCACTCCTCCAGCCCGCCTCGCCAAGACCCCGTCCTCAA 3704
E14fr TGGCCCTGCCAAAGAGGAGCTCTACTCCACCTCCAAGGATGCAAAAAAGTTCCTCCTCTG 3989
       **  **  **    ** **  * ** ** **  *  *  *    ** *   * *****
```

```
E14hs GCTCCTCCCAGACCTCGCCCGCCTCCCAGCCCCTGCCCAGAAAGCGCCCCCCGGTCACCC 3830
E14mm GCTCTTCACAGACCTCTCCAGCATCCCAGCCCCTGCCTAGGCGGTCCCCTCTGGCCACTC 3764
E14fr GTTCATCGCAAACCTCGACTCCATCCAAACAGAAG--AAGACAACATCCCTGAGTGGAAC 4047
      * ** ** ** *****  *  * *** * *    *   **       **    *     *

E14hs AGGCTGCTGGGGCCCT-------GCCCGGCCCCGGAG---CCTCC-CCGGTGCCCAAAAC 3879
E14mm CCACAGGAGGACCTCT-------GCCTGGCCCTGGGG---GGTCC-CTGGTGCCCAAGTC 3813
E14fr GAATAAAAACATCCCTAAAAAGGGCGTTACGCCGACGAACGGTCCACCAGCTGCTGAAAG 4107
                * **       **    * * *  *     *** *  *   *  *

E14hs GCCGGCGCGCACCCTTCTG---------GCGAAGC----AGCAC--AAGACGCAGAGATC 3924
E14mm ACCAGCACGGGCCCTTCTG---------GCTAAGC----AACAC--AAGACCCAGAAGTC 3858
E14fr AACAGCAGGTGCAGCCTTGAACTCAGATGAGAAGCCCTCAACACCAAAAACTCAGAAGTC 4167
       * **  *  *     **          *  ****    * ***  ** ** ****  **

E14hs GCCCGTGCGGATCCCGTTCATGCAGAGGCCGGCCCGGCGTGGGCCGCCACCGCTGGCTCG 3984
E14mm ACCTGTGCGGATCCCATTCATGCAAAGGCCAGCCAGGCGAGTGCCACCTCCACTGGCCAG 3918
E14fr ACCCGTCCGAATCCCTTTCATGCAGAATCCAGTCAAACCCAGACCCCTGTCACCCCTGGT 4227
       ** ** ** ***** ******** *  ** * *   *     ** *   * *

E14hs GGCAGTCCCGGAGCCGGGCCCCAGGGGCCGGGCGGGGACCGAGGCGGGCCCGGGGGCGCG 4044
E14mm GCCATCCCCAGAACCTGGCTCTAGGGGCCGAGCTGGGGCTGAGGGTACTCCTGGGGCACG 3978
E14fr TACAAACCAAACAGCTGGC--AAACCAGCAAATATGGTCCATGGGAAAGT-GGTGTCGCC 4284
       **  **      * ***   *     *      ** *   **        * * * *

E14hs CGGGGGCCGCCTGGGCCTGGTGCGTGTGGCCTCAGCCCTCTCCAGCGGCAGCGAGTCCTC 4104
E14mm TGGCAGCCGCCTGGGCCTGGTGCGTATGGCTTCAGCTCGCTCCAGTGGCAGCGAGTCCTC 4038
E14fr TGCCAGTCGATTAGAACTGCTCCGGATGACCTCAGCT---------GGTCGTGAGTCC-- 4333
       *   * **  * *  *** * **  ** * *****          **  * ******

E14hs CGACCGCTCGGGCTTCCGGCGACAGCTAACCTTCATCAAGGAGTCGCCGGG---CTTGCG 4161
E14mm GGATCGCTCAGGCTTCCGAAGACAGCTGACTTTCATCAAGGAATCCCCAGGTCTCCTTCG 4098
E14fr -GATCGCAATGGATTTTTGAGACAGATGACCTTCATCAAGGAATCAAAGACCGTACAGAA 4392
       ** ***   ** **     ***** * ** ********** **
```

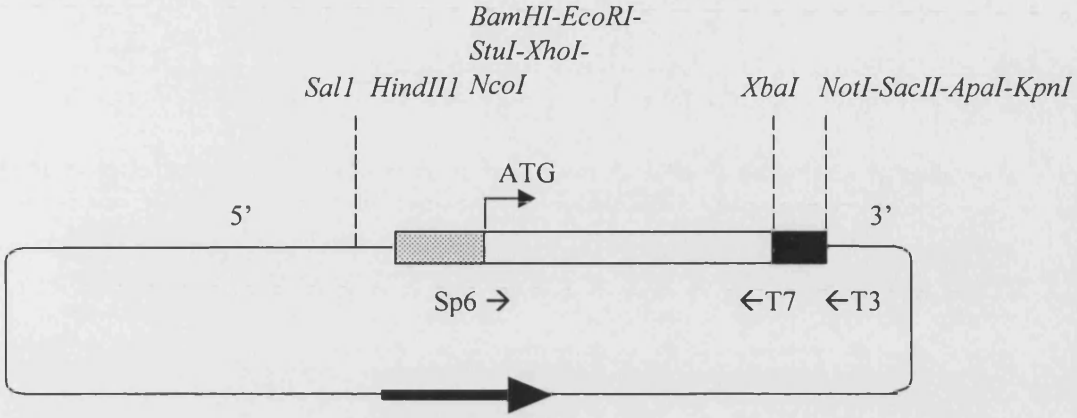
```
E14hs GCGCCGCCGCTCCGAGC-TGTCCTCGGCCGAGTCCGCGG------CCTCTGCCCCCCAGG 4214
E14mm GCGTCGCAGATCAGAGC-TGTCCTCTGCGGACTCCACGG------CCTCCACCTCCCAGG 4151
E14fr ACACGATAGTGCCAAATGCAGCATGTTCCGATCTCAAAAAAGACCCCTCCACCCTACCGG 4452
       *      *  *  *      * *   * **   *          ****  **   * **

E14hs GCGCCTCGCCCCGCCGCGGCCGGCCCGCGCTGCCCGCCGTCT----TCCTCTGC-----T 4265
E14mm CTGCTTCGCCCCGCCGTGGACGGCCTGCACTCCCTGCTGTCT----TTCTCTGC-----T 4202
E14fr CTCGGGAGCCGCGGCTGTTTTCCTTTGTTCTTCACGTTGTCAGGAACTCAAAGCAGCAGT 4512
            *** ** *           *  ** *  *  ***       *   **     *

E14hs CCTCGCGCTGCGAAGAGCTC-----------CGAGCGGCACCCCGGCAGGGC-------C 4307
E14mm CCTCTCGTTGCGATGAGCTG-----------CGGGTATCCCCACGGCAGCCC-------C 4244
E14fr TCAAACCCAGAGAAGAACACAAGTAAAAGATCAAGGACAGCCACAACGGGTCAAGAGACC 4572
       *   *   * ** ** *             *  *     ** *  * *  *       *

E14hs CGGCCCCGGCC--CGGCAGCGGCCCCCCGCGGCCCGACCCAGCCCTGGCGAGCG------ 4359
E14mm TGGCAGCA--------CAGAGGTCCCCACAGGCCAAGCCAGGTCTCGCCCCACG------ 4290
E14fr TGATCCTGGCCTTCAGCAGCAGGCAACACTCTCCAGAGCAACCTCCAGCGAAAGGTACAG 4632
       *             ***  * *  *     **    *         *    *

E14hs ------------CCCTGCCCGGCGCACCACCTCCGAGAGCCCGTCCCGCCTGCCTGTGC- 4406
E14mm ------------TGCGCCCAGACGCACCAGCTCTGAGAGCCCCTCACGCCTGCCTGTGC- 4337
E14fr AAACACAAGAGGTTTAAGCAGAAGAACTAGCTCTGAAAGCCCATGCAGGCTGGCAAAGCA 4692
                     * *  * ** * *** ** ***** *   * *** *   **

E14hs -------GCGCGCCCGCCGCCCGGCCGGAGACTG--------------------TCAAGCG 4440
E14mm -------GGGCGAGCCCTGGACGGCCTGAGACAG--------------------TCAAGCG 4371
E14fr AAGCAAAGCTGGCACGGTGTCTGGCGTCAGGCAGCAACAAGACAAGGATACCTTCAAACG 4752
            *   *  *   *   ***   ** * *                    **** **

E14hs CTACGCGTCGCTGCCGCACATCAGCGTGGCCCGCAGGCCCGACGGCGCCGTCCCCGCGGC 4500
E14mm GTACGCATCCCTGCCACATATTAGTGTGTCCCGCAGGTCCGATAGCGCTGTCTCTGTGCC 4431
E14fr TCATGCCTCATCACCAAGCATCAACATACTGAGCCGCGTCACCAGCCGCTCTTCCCTCCG 4812
       * ** **    **    ** *   *     ** *   *    **       *
```

```
E14hs CCCTGCCTCAGCCGACGCCGCGCGCCG---CAGCAGCGACGGGGAGCCCCGGCCGC-TCC 4556
E14mm CACCACCCAGGCCAATGCCACTCGCCG---GGGAAGTGATGGCGAGGCCAGGCCGC-TGC 4487
E14fr CTCATCATCGTCTGATTCAAGTAGCAGAGCCAAGAGTGAGGACGAGACAAAGAAGTATGG 4872
      * *  *     *  *  *     ** *       ** ** *  *** *   *  *  *

E14hs CCAGGG----TGGCCGCGCCG-GGCACGACCTGGCGGCGCATCCGAGATGAGGACGTGCC 4611
E14mm CCAGGG----TAGCTCCTCCG-GGTACGACCTGGCGTCGAATCAAAGATGAAGATGTCCC 4542
E14fr ACAGAAATCCCGGGTACTTGACAGGGCCACCTGGAGGAGGATCAGGGATGAAGATGTTCC 4932
       ***        *   *      *  * ****** *  * ***   ***** ** ** **

E14hs CCACATCCTGCGCAGCACGCTTCCCGCCACGGCCCTGCCACTGCGGGGCTC--------- 4662
E14mm CCACATCCTGCGCAGCACACTGCCTGCCACTGCCCTGCCTCTCAGGGTCTC--------- 4593
E14fr TCACATCTTAAAAAGCACCCTCCCGGCTAATGCCTTACCCCTGGTGGCTTCTCCCGAGGG 4992
       ****** *    ***** ** ** ** *  *** * ** **   **  **

E14hs --CACGCCCGAGGACG---CCCCGGCCGGG-------------CCCCCGCCG------CG 4698
E14mm --GTCACCCGAAGACA---GCCCCGCTGGA-------------ACTCCACAG------CG 4629
E14fr GGACCAGCCGAAGCTGCAGGCTCCTTTGGGTAAGTTGCCAACCATTCTGCTGGCCTCCCG 5052
         *  **** *       * *    **                 *  * *      **

E14hs CAAGACCAGCGACGCCGTGGTCCAGACCGAGGAGGTCGCCGCCCCCAAGACCAACTCCAG 4758
E14mm CAAGACCAGTGACGCAGTGGTGCAAACAGAGGACGTGGCCACTTCTAAGACCAACTCTAG 4689
E14fr CAAGACAAGTGATGCAACGGTCCAGACCGAGGACTTTTCCAAC---AAGATCAGCTCCAG 5109
      ****** ** ** **   *** ** ** *****  *  **      **** ** *** **

E14hs CACGTCCCCGA------------GCCTGGAGACC----AGGGAGCCCCCCGGG--GCCCC 4800
E14mm CACGTCACCTA------------GCCTGGAGAGC----AGGGATCCGCCACAG--GCCCC 4731
E14fr CACCTCTCCAACAGTGGAAGTTGGCCCAGAGATCGCAGAAGAAACGGTCCGACCCGCACC 5169
      *** ** ** *           ***  **** *    * * * *   *      ** **

E14hs CGCCGGCGGCCAGCTCTCCCTCCTCGGCAGCGACGTGGACGGTCCCAGCCTCGCCAAGGC 4860
E14mm TGCCAGCGGCCCTGTGGCTCCCCAGGGCAGCGATGTGGATGGACCAGTACTCACCAAGCC 4791
E14fr GCTCAGAAATGAAGGTACGACCTCCGGAAATAACCTCCA-GGATGGAGACTCAGATTGCT 5228
        * *          *   *   ** *   *  *  * **       ***     *
```

```
E14hs TCCC-----ATCTCCGCACCCTTCGTGCACGAGGGCCTGGGGGTCGCCGTGGGGGGCTTC 4915
E14mm TCCT-----GCCTCAGCGCCCTTCCCCCATGAGGGTCTGAGTGCTGTCATAGCAGGCTTT 4846
E14fr TGTTAAAAAGCCTCAGCAACGCTTCCATGGGAACGCCAGACAACCACGCTGGTGGGAGTG 5288
      *          *** **  *  *       **  * * *          * *  **  *

E14hs --CCCGCCAG---CCGGCACGGCTCCCCCAGCCGCTCGGCCCGAGTACCCCCCTTCAACT 4970
E14mm --CCCACCAG---CAGGCATGGCTCCCCCAGCAGGGCTGCACGGGTTCCCCCCTTCAACT 4901
E14fr GGCCGGTCTATTTCCGCCAGGGAACTCCAAGCAAGTCTGCCAGAATCACTCCGTTTAATT 5348
        **   *     * * ** **  * ** ***    * **  *  *  * ** ** ** *

E14hs ATGTGCCCAGCCCCATGGT---GGTCGCAGCCACCACCGACTCGGCCGCGGAGAAAGCCC 5027
E14mm ATGTGCCCAGCCCTATGGCAGCGGCCACAATGGCCAGTGACTCCGCAGTGGAGAAAGCCC 4961
E14fr ACACACCCAACCCACTGGCCTGCAGTAAAAGCGCACAGAACCAGGCAGCCAAAACCAATG 5408
      *    **** ***  ***          *    *     **   ** *   * *

E14hs CGGCCACTGCCTCCGCCACCCTCCTGGAATAG 5059
E14mm CGGTCTCCTCTCCAGCCAGTCTCCTGGAGTAG 4993
E14fr AAAAGCAGGCGGAGGGCAGGGAGTCG------ 5359
               *    * **       *
```

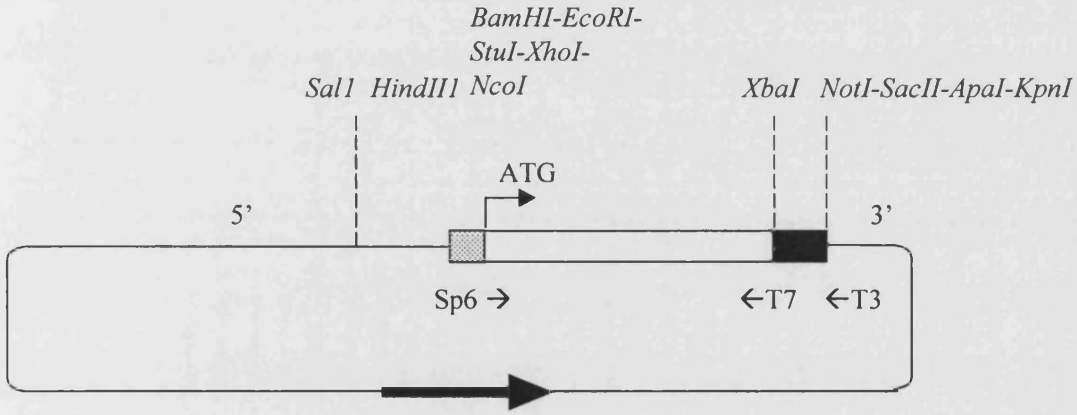
Appendix, Section 4A

mApc1 CAT reporter construct and CAT plasmids



p*CAT*- control

Figure. Plasmid map of p*CAT*-controlThe dotted square indicates the SV40 promoter, the open square shows the *CAT* gene and the black square indicates the SV40polyA. The ampicillin resistance gene is indicated by the black arrow.



p*CAT*-basic

Figure. Plasmid map of p*CAT*-basicThe dotted square indicates the minimal SV40 promoter,the open square shows the *CAT* gene and the black square indicates the SV40polyA. The ampicillin resistance gene is indicated by the black arrow.

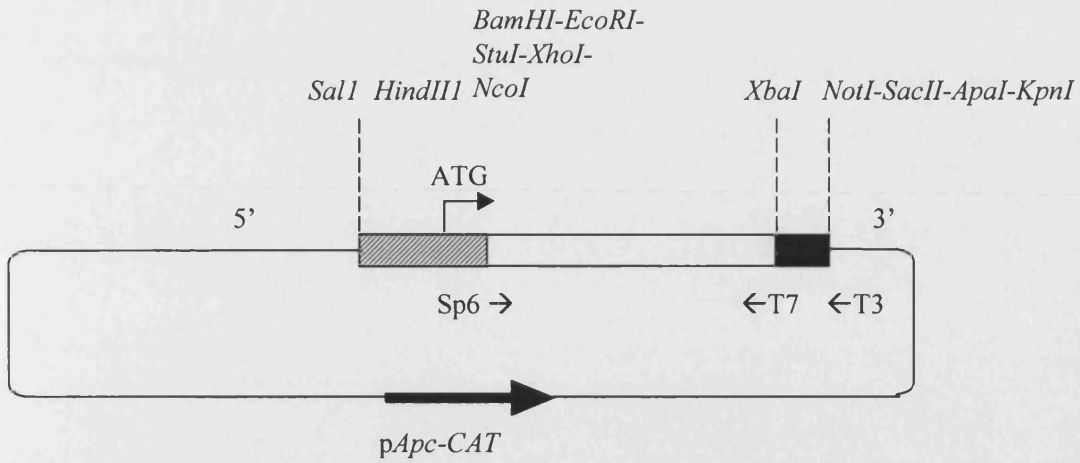
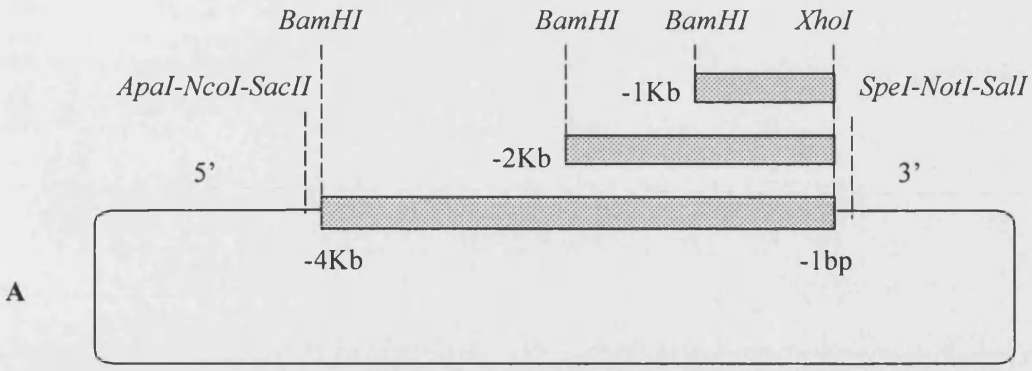


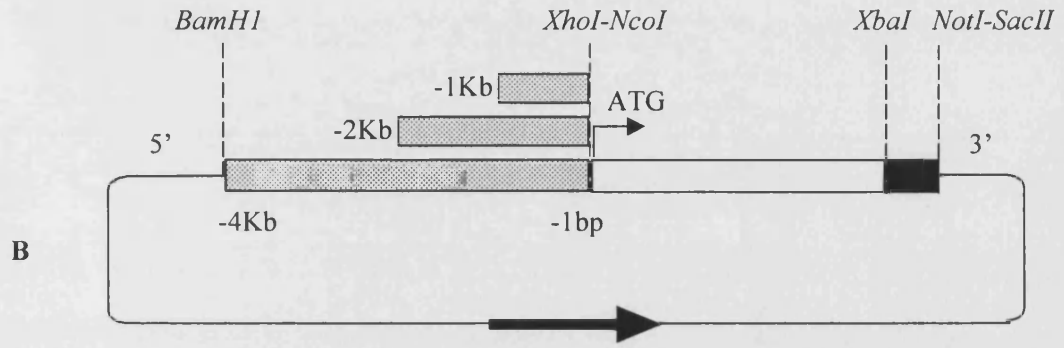
Figure. Plasmid map of p*Apc-CAT*. The dotted square indicates the *Apc* promoter (-290 to +317bp), the open square shows the *CAT* gene and the black square indicates the SV40polyA. The ampicillin resistance gene is indicated by the black arrow.

Appendix, Section 4B

frApc1 GFP reporter constructs



frApc1 PCR fragments- pGEM-T



frApc1 deletion constructs- pCS2:GFP^{-CMV}

Figure. Plasmid map of *frApc1* deletion constructs- pCS2:GFP^{-CMV}. (A). Shows the cloning of the frCdx1 PCR products into pGEM-T. The dotted square indicates the 4.0, 2.0 and 1.0 Kb fr*Apc1* PCR products. (B) Cloning of the fr*Cdx1* PCR products into the pCS2:GFP$_{-CMV}$ vector using the *BamHI/ XhoI* sites. The open square shows the *GFP* gene and the black square indicates the SV40polyA. The ampicillin resistance gene is indicated by the black arrow.

Appendix, section 5A

frCdx1 GFP reporter construct, microinjection into the zebrafish, 80% epiboly

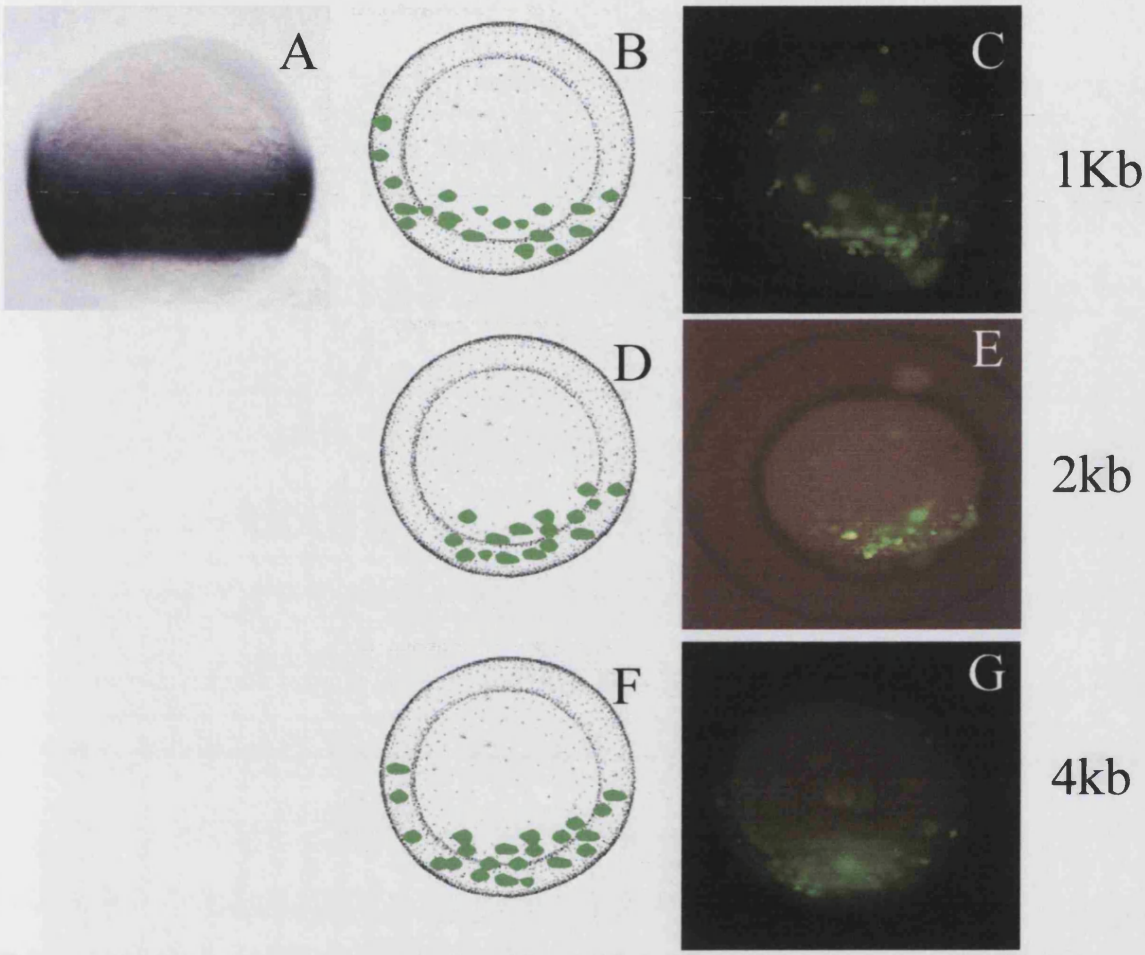


Figure. Transient expression analysis of the *frCdx1* GFP constructs in to the zebrafish embryo at 80% epiboly. **A**. Lateral view of the *Zcad* expression at 50% epiboly. **B**,**D** and **F**. Expression map representing the GFP expression pattern shown for each *frCdx1* reporter construct. **C**,**E** and **G**. GFP expression of each construct in the zebrafish embryo at 80% epiboly (top view).

***frCdx1* GFP reporter construct, microinjection into the zebrafish, 1 somite stage**

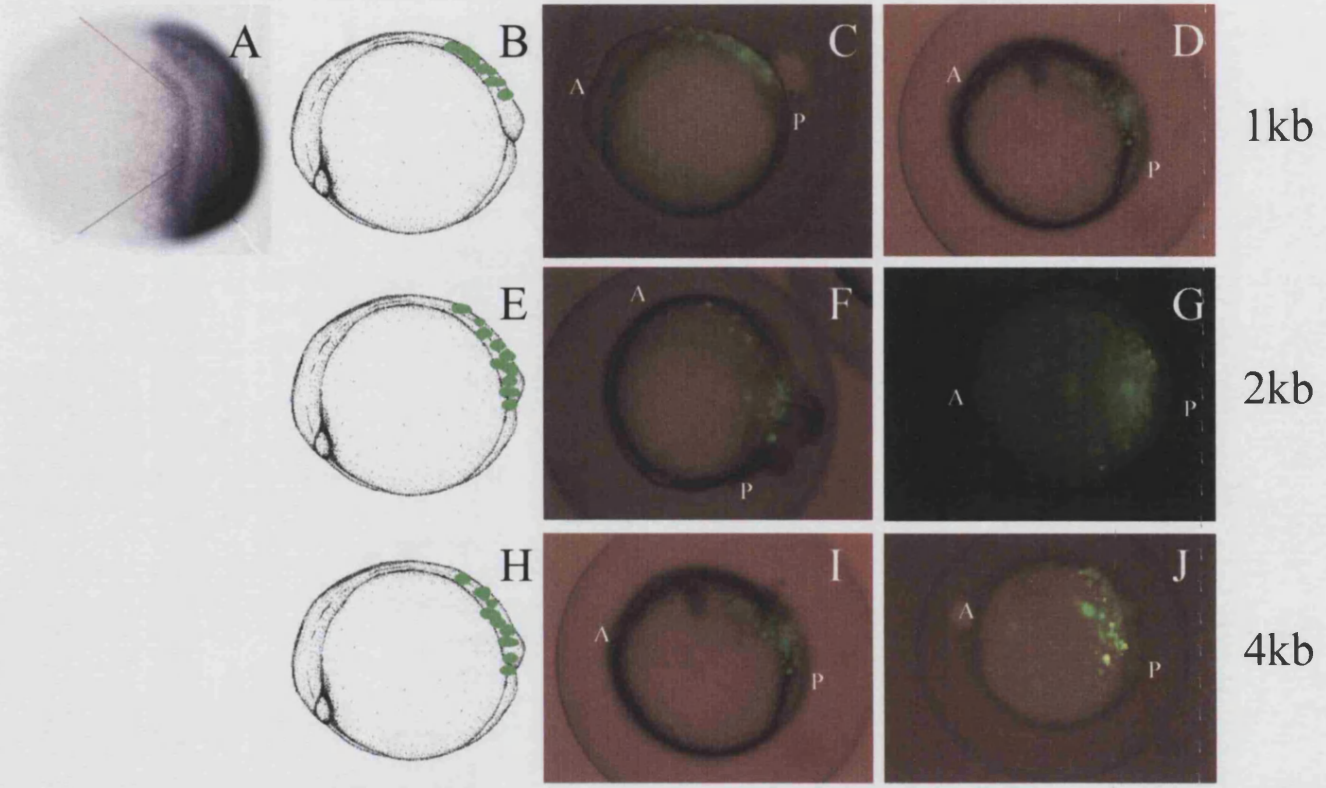


Figure. Transient expression analysis of the *frCdx1* GFP constructs in to the zebrafish embryo at 1 somite stage. **A**. Lateral view of the *Zcad* expression at bud stage.**B**,**E** and **H**. Expression map representing the GFP expression pattern shown for each *frCdx1* reporter construct. **C** and **D**. GFP expression of the –1.0Kb *frCdx1* GFP construct. **F** and **G.** GFP expression of the –2.0Kb *frCdx1* GFP construct. **I** and **J** GFP expression of the –4.0Kb *frCdx1* GFP construct in the zebrafish embryo at bud stage (lateral view). (A) anterior, (P) posterior.

frCdx1 GFP reporter construct, microinjection into the zebrafish, 14 somites stage

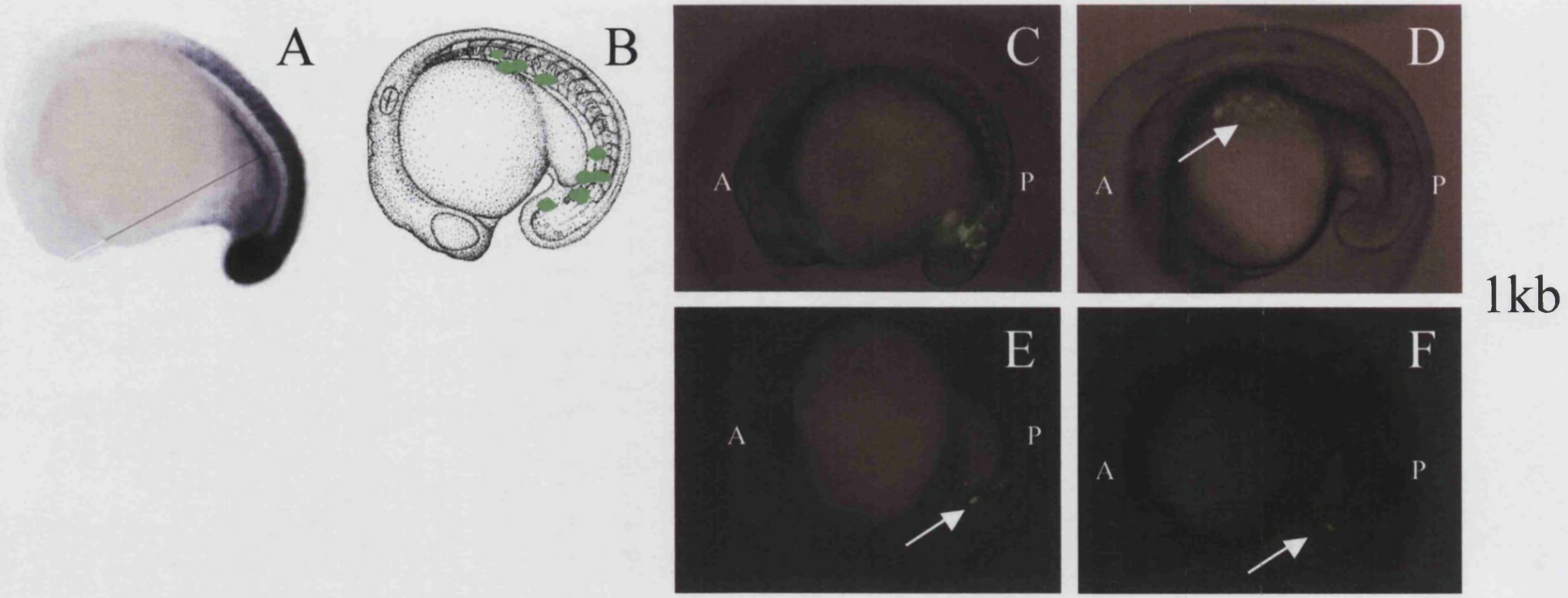


Figure. Transient expression analysis of the 1.0Kb *frCdx1* GFP construct in to the zebrafish embryo at 14 somites stage. **A**. Lateral view of the *Zcad* expression at 18 somite stage. **B**. Expression map representing the GFP expression pattern shown by the –1.0Kb *frCdx1* GFP construct. **C- F**. GFP expression in the tail bud and ventral mesenchyme region in the zebrafish embryo at 18 somite stage (lateral view). Embryos orientated (A) anterior to (P) posterior.

frCdx1 GFP reporter construct, microinjection into the zebrafish, 14 somites stage

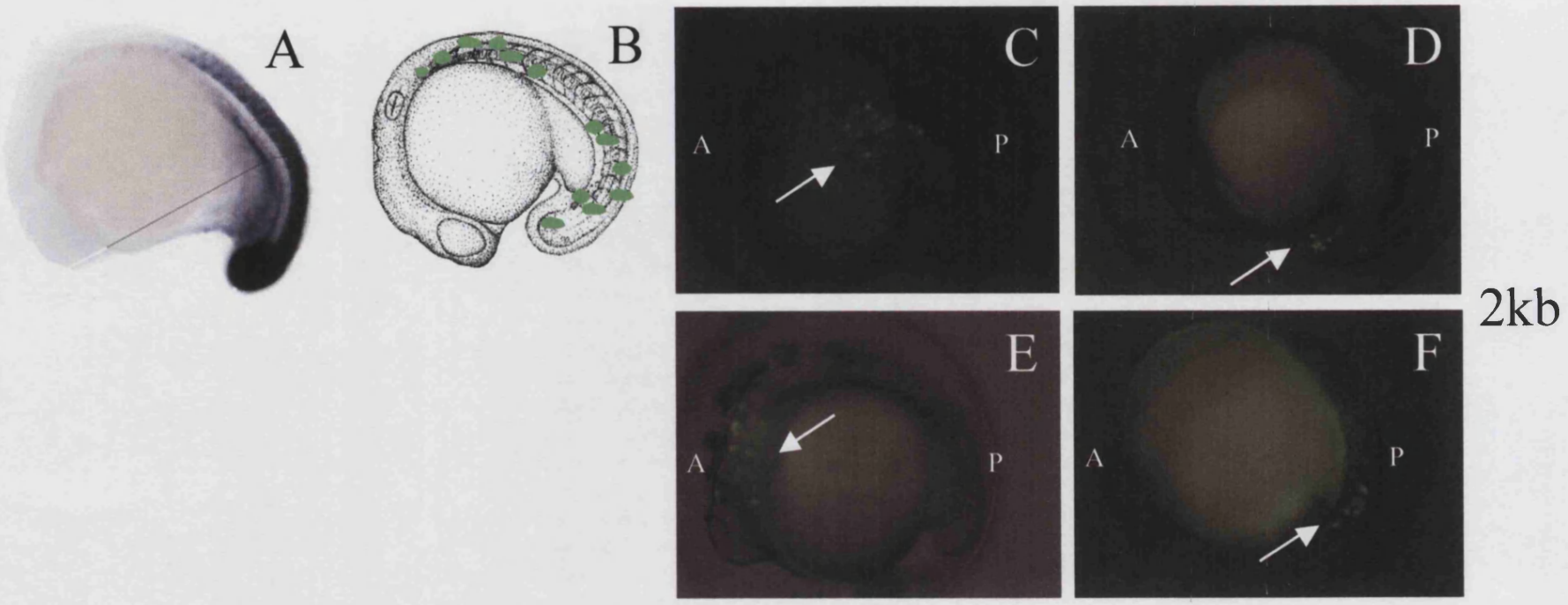


Figure. Transient expression analysis of the 2.0Kb *frCdx1* GFP construct in to the zebrafish embryo at 14 somites stage. **A**. Lateral view of the *Zcad* expression at 18 somite stage. **B**. Expression map representing the GFP expression pattern shown by the –2.0Kb *frCdx1* GFP construct. **C- F**. GFP expression in the tail bud, ventral mesenchyme and spinal chord region in the zebrafish embryo at 18 somite stage (lateral view). Embryos orientated (A) anterior to (P) posterior.

frCdx1 GFP reporter construct, microinjection into the zebrafish, 14 somites stage

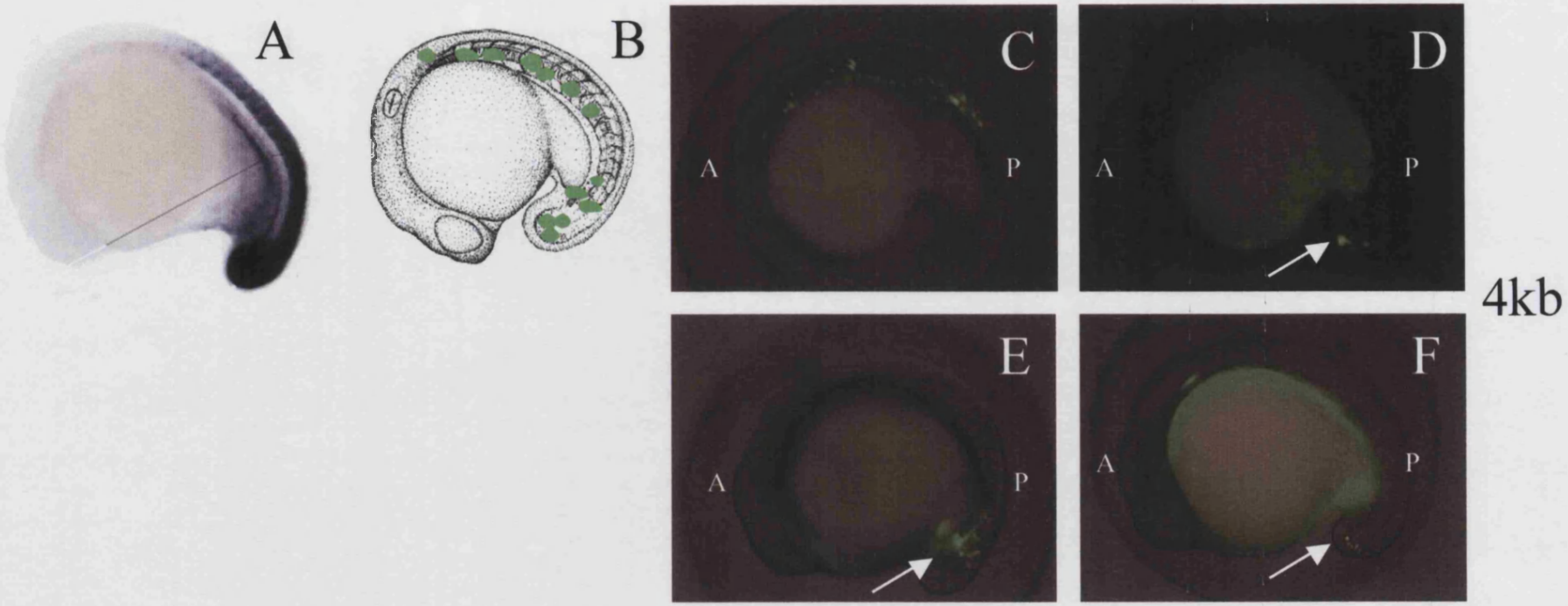


Figure. Transient expression analysis of the 4.0Kb *frCdx1* GFP construct in to the zebrafish embryo at 14 somites stage. **A**. Lateral view of the *Zcad* expression at 18 somite stage. **B**. Expression map representing the GFP expression pattern shown by the –4.0Kb *frCdx1* GFP construct. **C- F**. GFP expression in the tail bud and spinal chord region in the zebrafish embryo at 18 somite stage (lateral view). Embryos orientated (A) anterior to (P) posterior.

frCdx1 GFP reporter construct, microinjection into the zebrafish, 24hpf

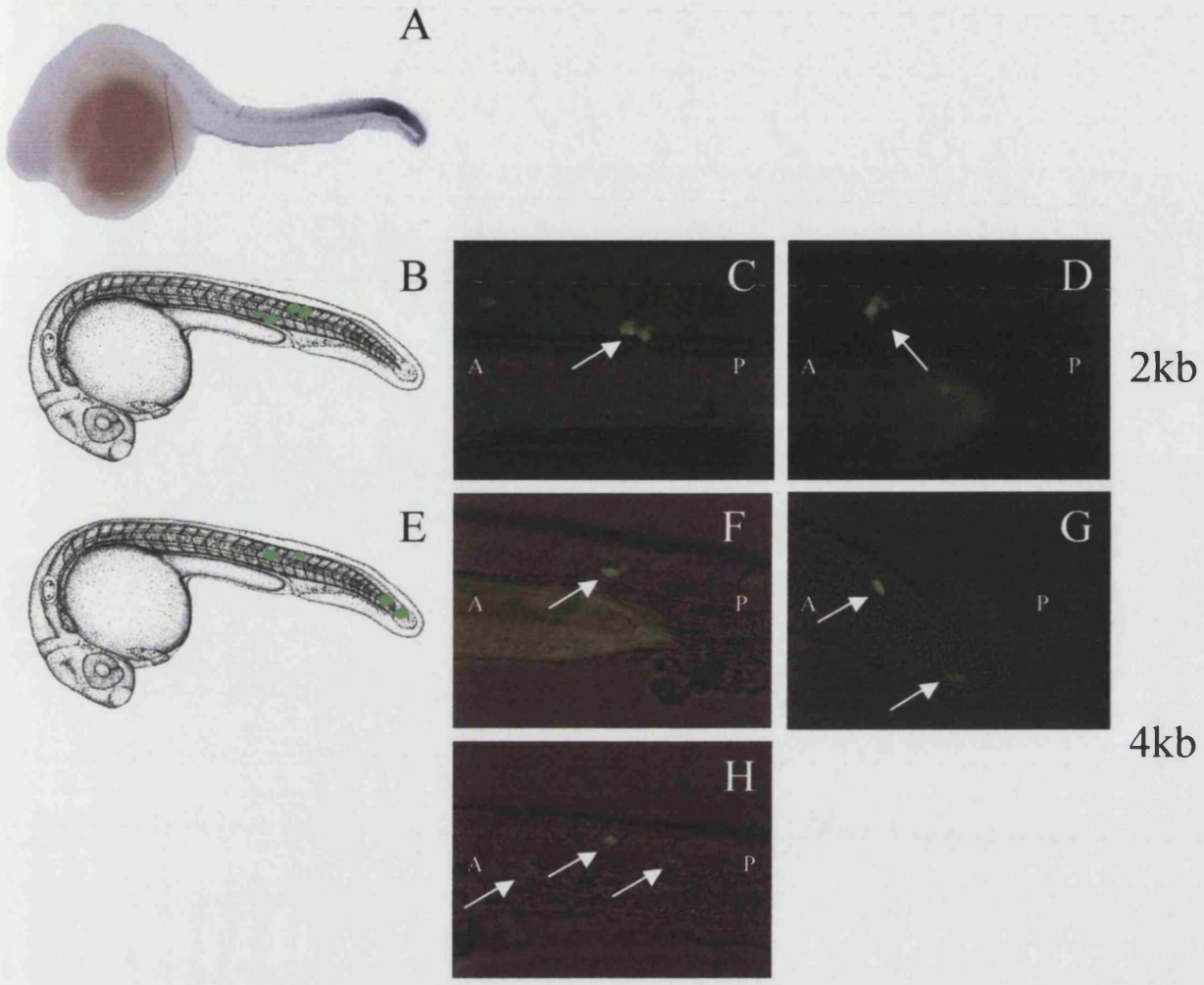


Figure. Transient expression analysis of the *frCdx1* GFP constructs in to the zebrafish embryo at 25hpf. **A**. Lateral view of the *Zcad* expression at 25hpf. **B** and **E**. Expression map representing the GFP expression pattern shown for the –2.0Kb *frCdx1* GFP and –4.0Kb *frCdx1* GFP reporter constructs. **C** and **D**. GFP expression of the –2.0Kb *frCdx1* GFP construct in the spinal cord region. **F**, **G** and **H**. GFP expression of the –2.0Kb *frCdx1* GFP construct in the spinal cord and tail bud region in the zebrafish embryo at 25hpf (lateral view). Embryos orientated (A) anterior to (P) posterior.